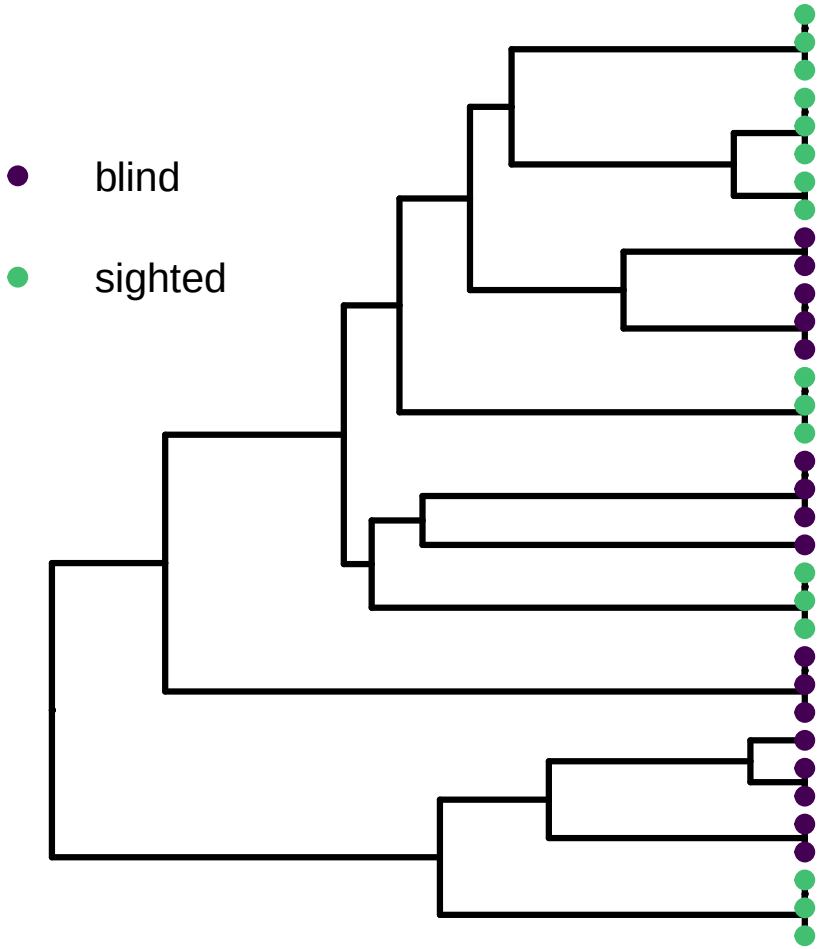
# A Phylogenetic Framework to Simulate Synthetic Inter-species RNA-Seq Data

Paul Bastide[1], Charlotte Soneson[2,3], David B. Stern[4,5], Olivier Lespinet[6], and Mélina Gallopin[6]

[1] IMAG, Université de Montpellier, CNRS
[2] Friedrich Miescher Institute for Biomedical Research, Basel
[3] SIB Swiss Institute of Bioinformatics, Basel
[4] Department of Integrative Biology, University of Wisconsin-Madison
[5] National Biodefense Analysis and Countermeasures Center (NBACC), Fort Detrick
[6] I2BC, Université Paris-Saclay, CEA, CNRS

StatOmique, October 2022

# Vision Loss in the Crayfish Familly (Stern and Crandall, 2018)




Stern and Crandall (2018)

*Orconectes australis*

*Cambarus dubius*

Evolution of gene expression underlying vision loss ?

# Outline

# Differential Expression Analysis for RNA-seq Data

## Gene expression matrix:

| | condition A | | | condition B | | |
|---|---|---|---|---|---|---|
| | r1 | r2 | r3 | r1 | r2 | r3 |
| gene 1 | ... | ... | ... | ... | ... | ... |
| gene 2 | ... | ... | ... | ... | ... | ... |
| ... | ... | ... | ... | .... | ... | ... |
| ... | ... | ... | ... | .... | ... | ... |

## Observations:

**counts** $y_{gi}$ : expression level for gene $g$ and sample $i$ $(p \times n)$

## Question:

Find genes $g$ differentially expressed between conditions A and B ?

# GLM Modelling Counts : DESeq2 (Love et al., 2014)

$$Y_{gi} \sim \text{NB}(\mu_{gi}, \alpha_g)$$
$$\mu_{gi} = s_i q_{gi}$$
$$\log_2(q_{gi}) = \mathbf{X}_i.\boldsymbol{\theta}_g$$

- $Y_{gi}$: counts ($p \times n$)
- $\alpha_g$ : dispersion
- $s_i$ : size factor
- $q_{gi}$ : proportional to expected true concentration
- $\mathbf{X}$ : design matrix ($n \times q$)
- $\boldsymbol{\theta}_g$ : $\log_2$ fold changes ($q \times 1$)

# Linear Model on Normalized Data : `limma` (Smyth, 2004)

$$\tilde{\mathbf{Y}}_g = \mathbf{X}\boldsymbol{\theta}_g + \mathbf{E}_g \qquad \mathbf{E}_g \sim \mathcal{N}(\mathbf{0}, \sigma_g^2 \mathbf{I}_n)$$

- $\tilde{\mathbf{Y}}_g$: normalized data for gene $g$ $(n \times 1)$
- $\sigma_g^2$: (moderated) gene specific variance

# Linear Model on Normalized Data : `limma` (Smyth, 2004)

$$\tilde{\mathbf{Y}}_g = \mathbf{X}\boldsymbol{\theta}_g + \mathbf{E}_g \qquad \mathbf{E}_g \sim \mathcal{N}(\mathbf{0}, \sigma_g^2 \mathbf{I}_n)$$

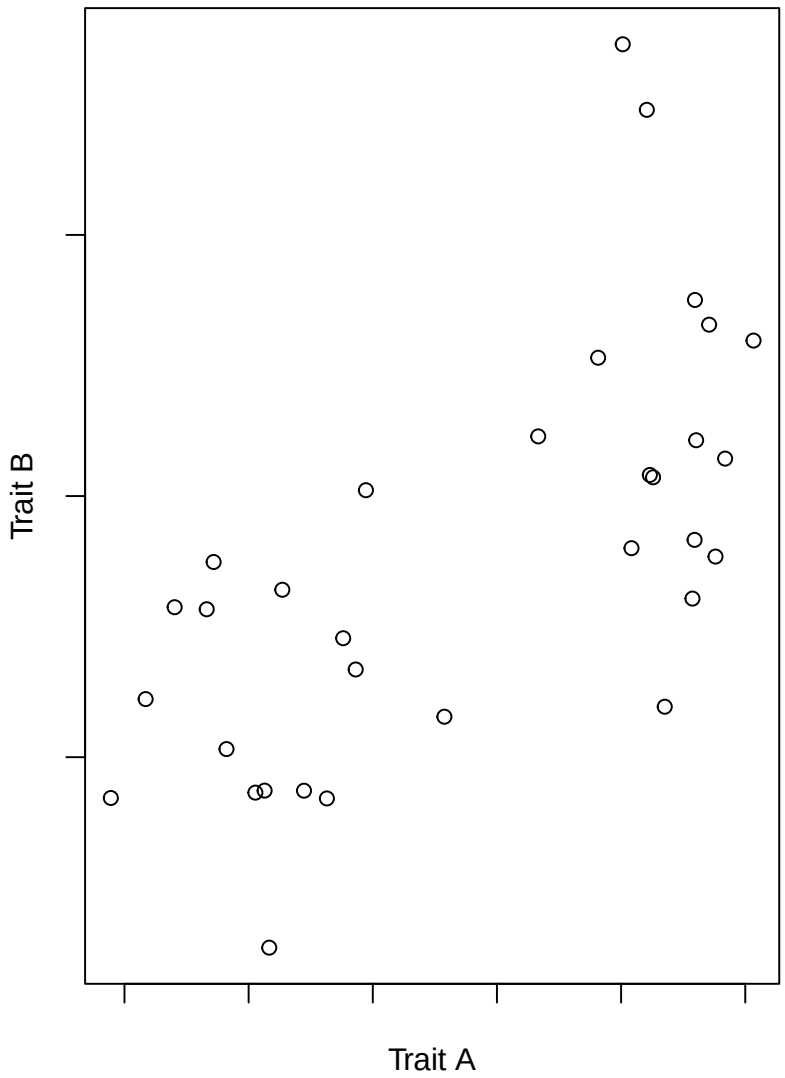
- $\tilde{\mathbf{Y}}_g$: normalized data for gene $g$ $(n \times 1)$
- $\sigma_g^2$: (moderated) gene specific variance

## $\log_2$ CPM

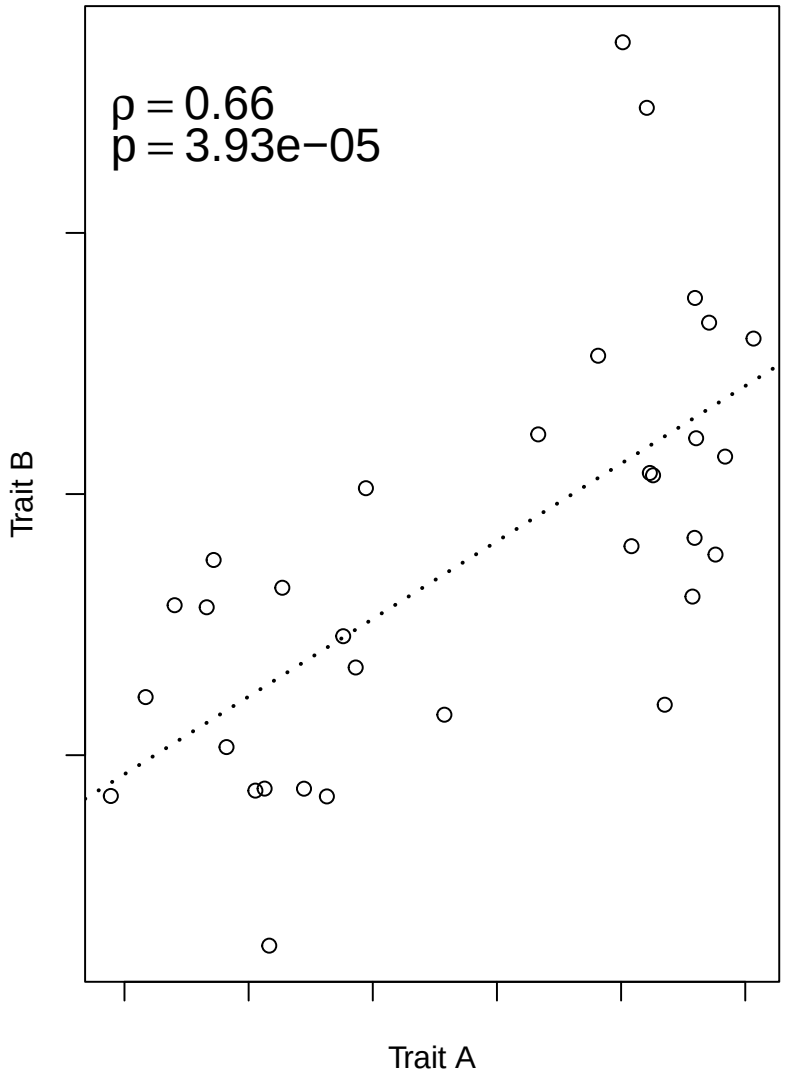
$$\tilde{Y}_{gi} = \log_2 \left[ \frac{Y_{gi} + 0.5}{M_i + 1} \times 10^6 \right] \qquad M_i = \sum_g Y_{gi} m_i$$

- $M_i$ the normalized library size
- $m_i$ a normalization factor [using eg RLE or TMM]

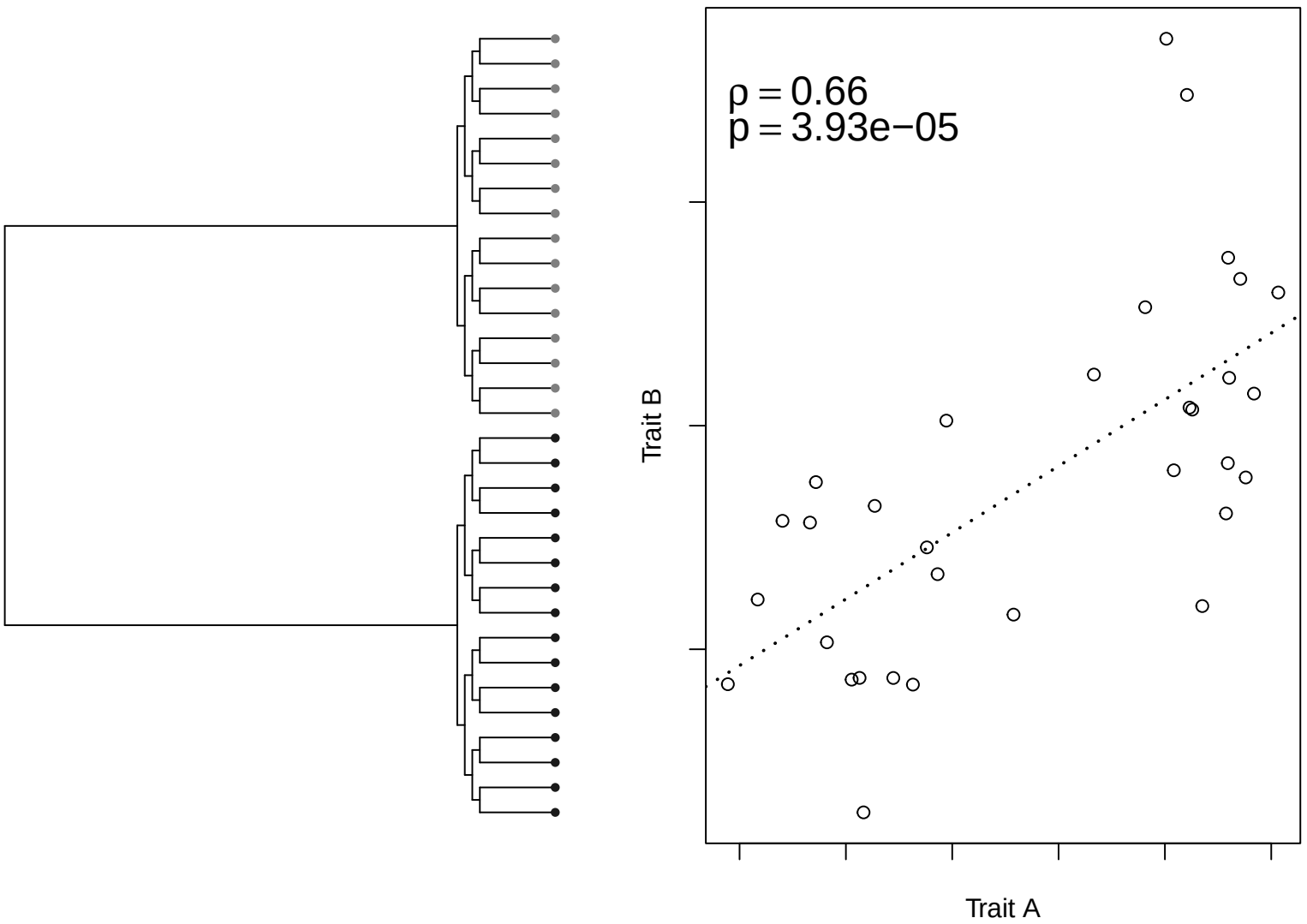
# Phylogenetic Comparative Methods



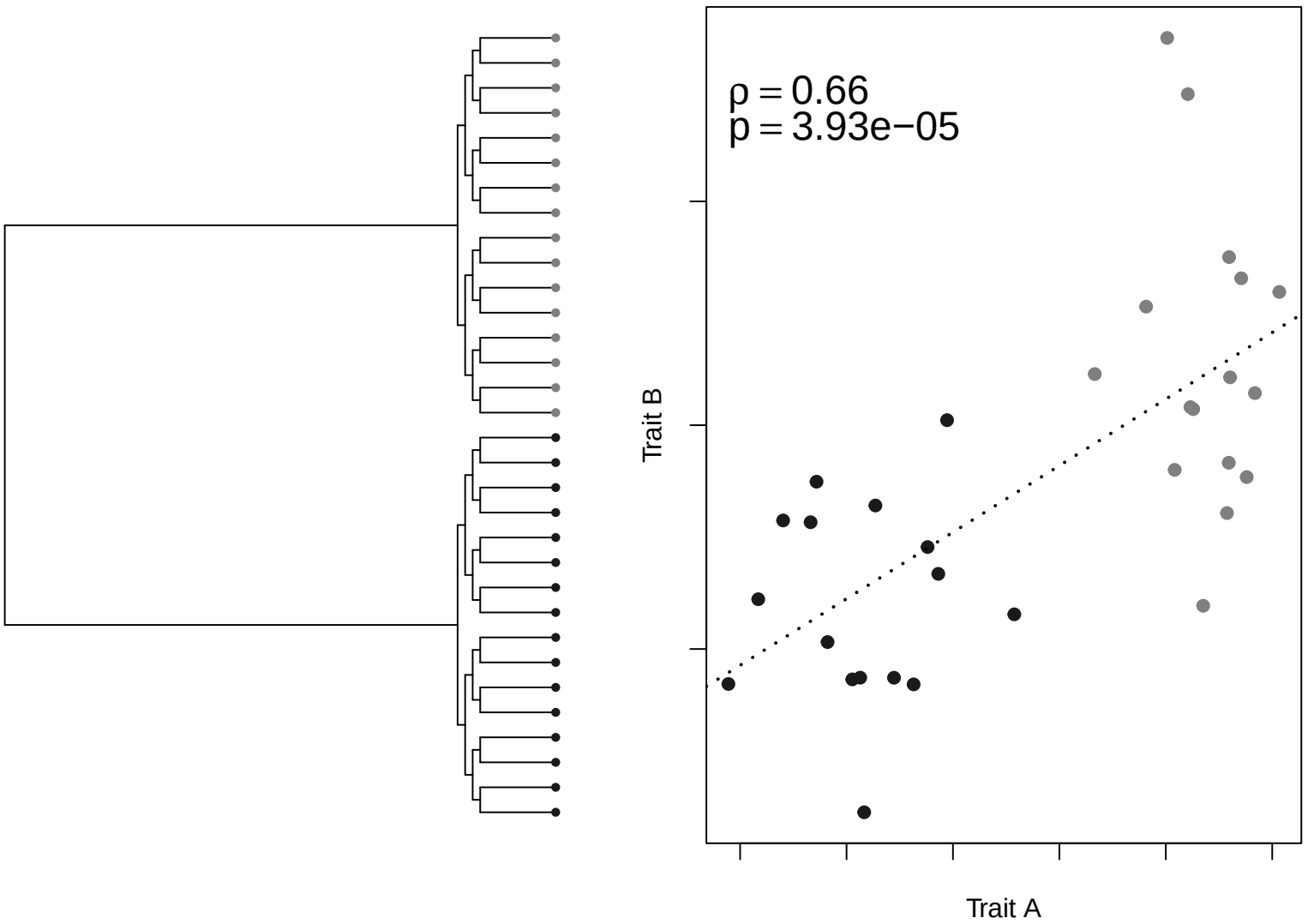
# Phylogenetic Comparative Methods



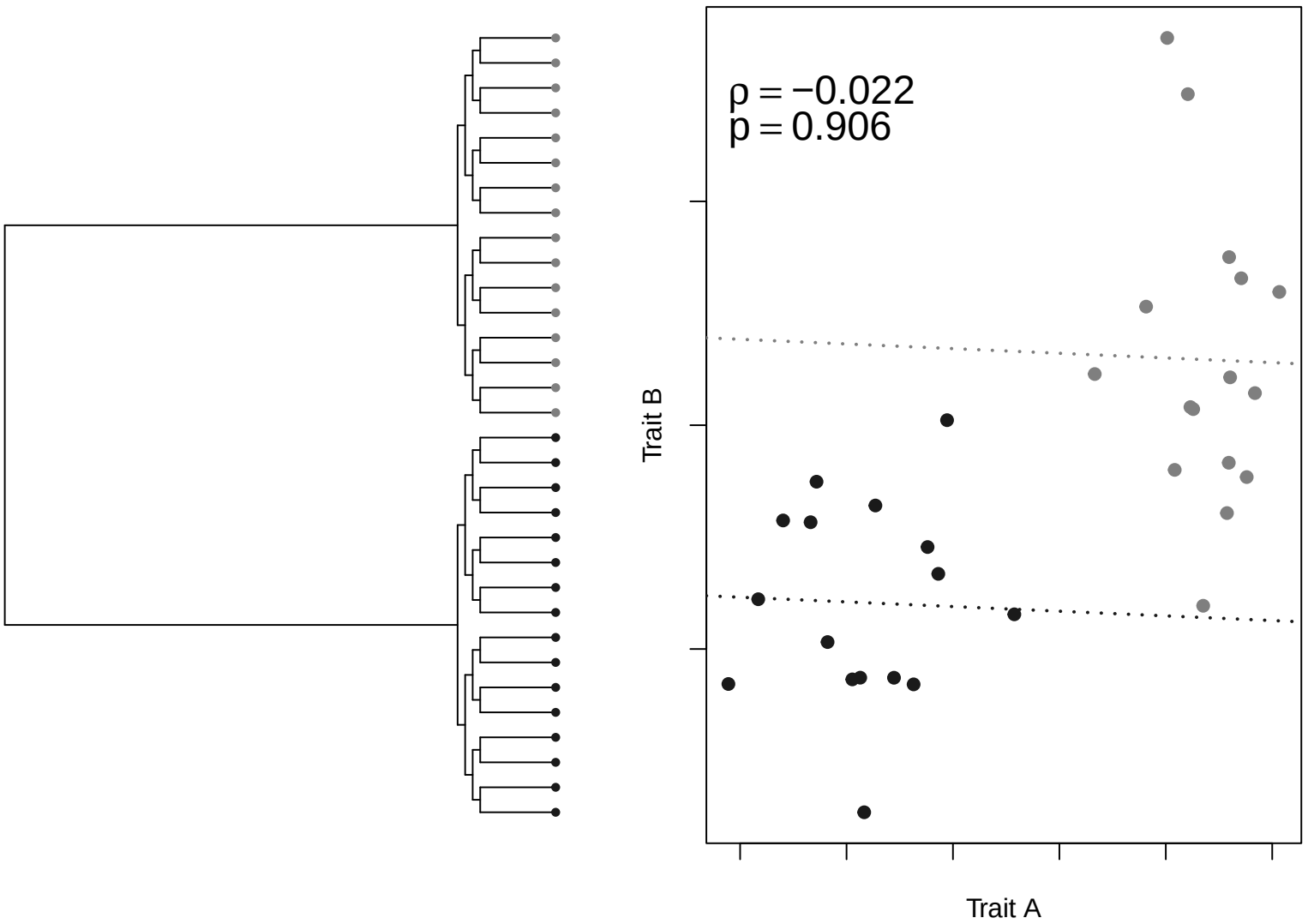
# Phylogenetic Comparative Methods



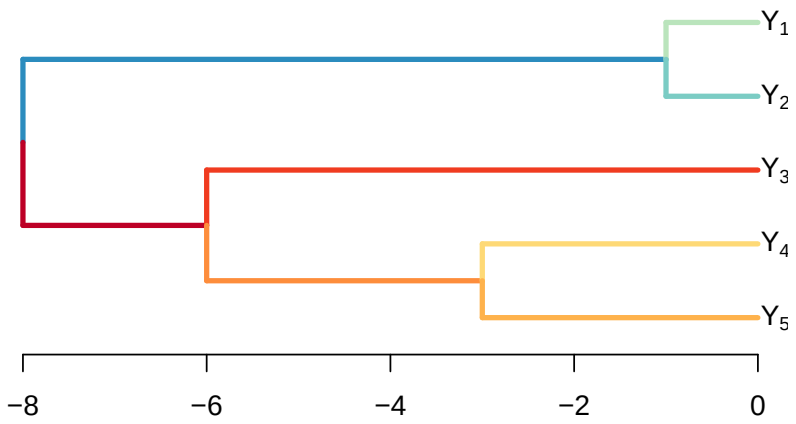
# Phylogenetic Comparative Methods

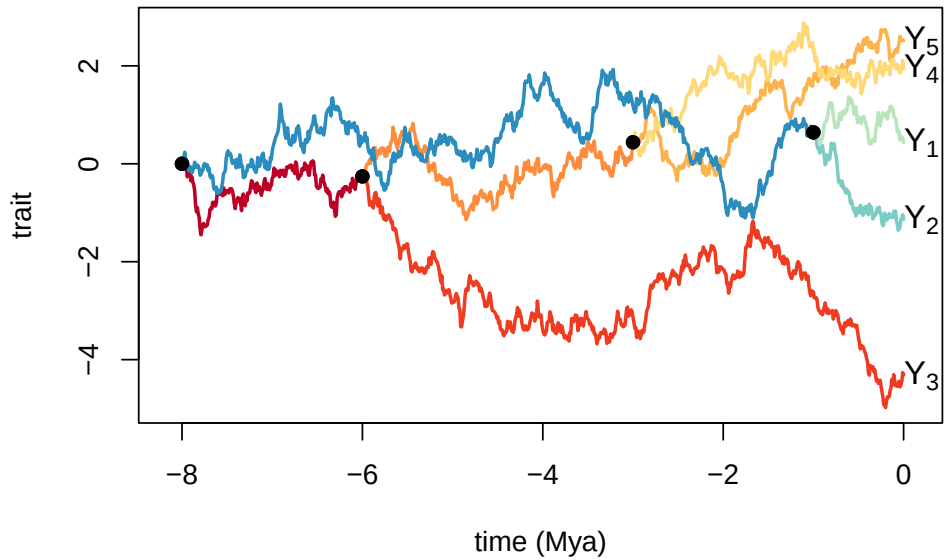


# Phylogenetic Comparative Methods



# Brownian Motion on a Tree (Felsenstein, 1985)

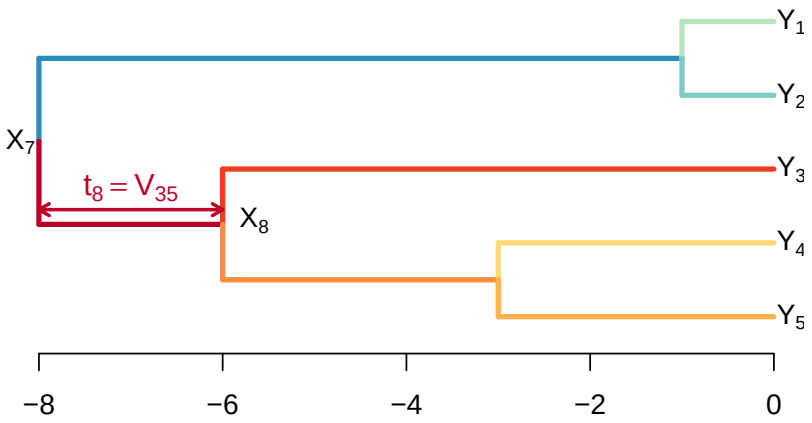


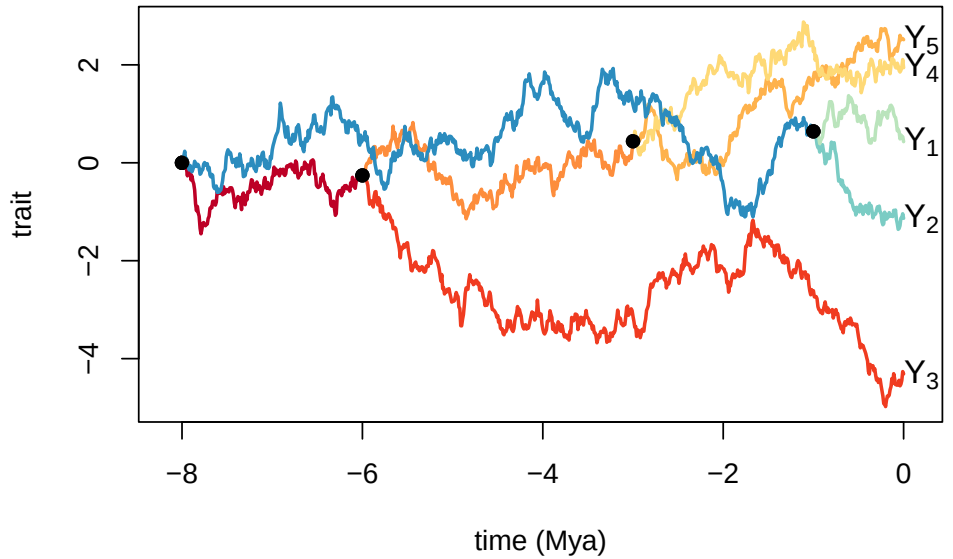




- The trait evolves like a BM in time
- Speciation $\rightarrow$ two independent processes
- Only tip values are measured

# Brownian Motion on a Tree

(Felsenstein, 1985)

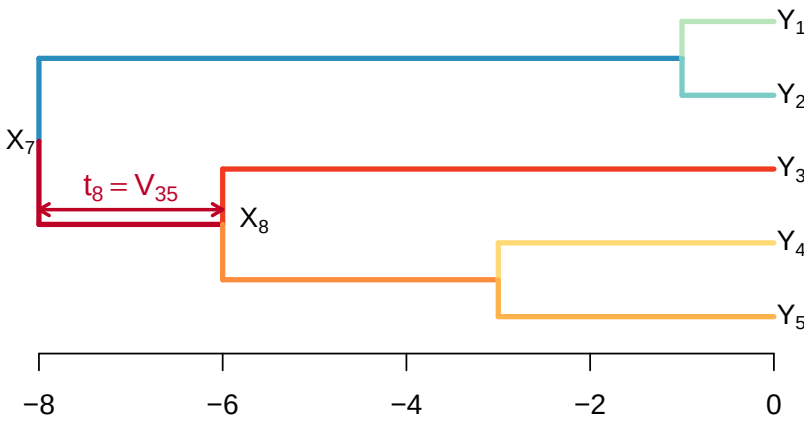


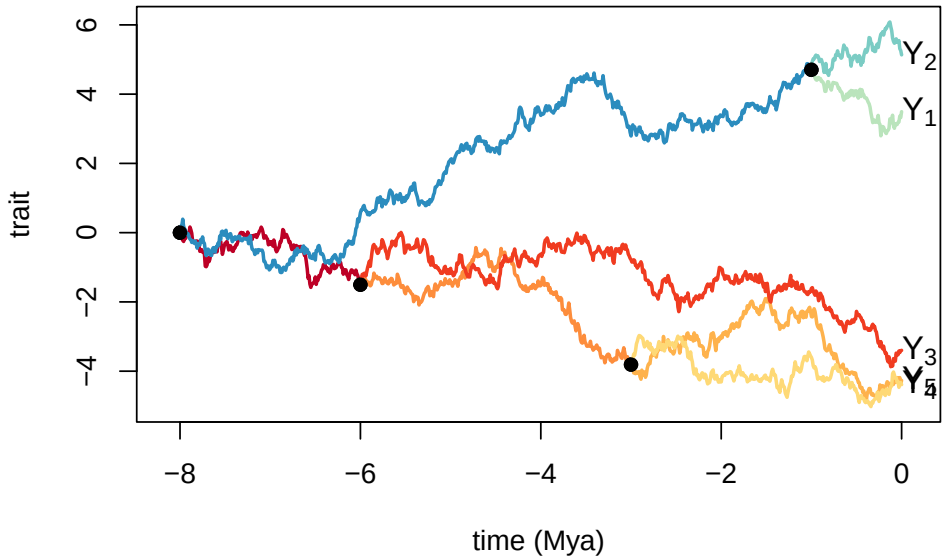




- SDE: $dX_t = \sigma dB_t$
- Covariances: $\mathbb{C}\text{ov}(Y_i, Y_j) = \sigma^2 V_{ij}$
- Distribution: $\mathbf{Y} \sim \mathcal{N}(\mu \mathbf{1}_n, \sigma^2 \mathbf{V})$

# Brownian Motion on a Tree

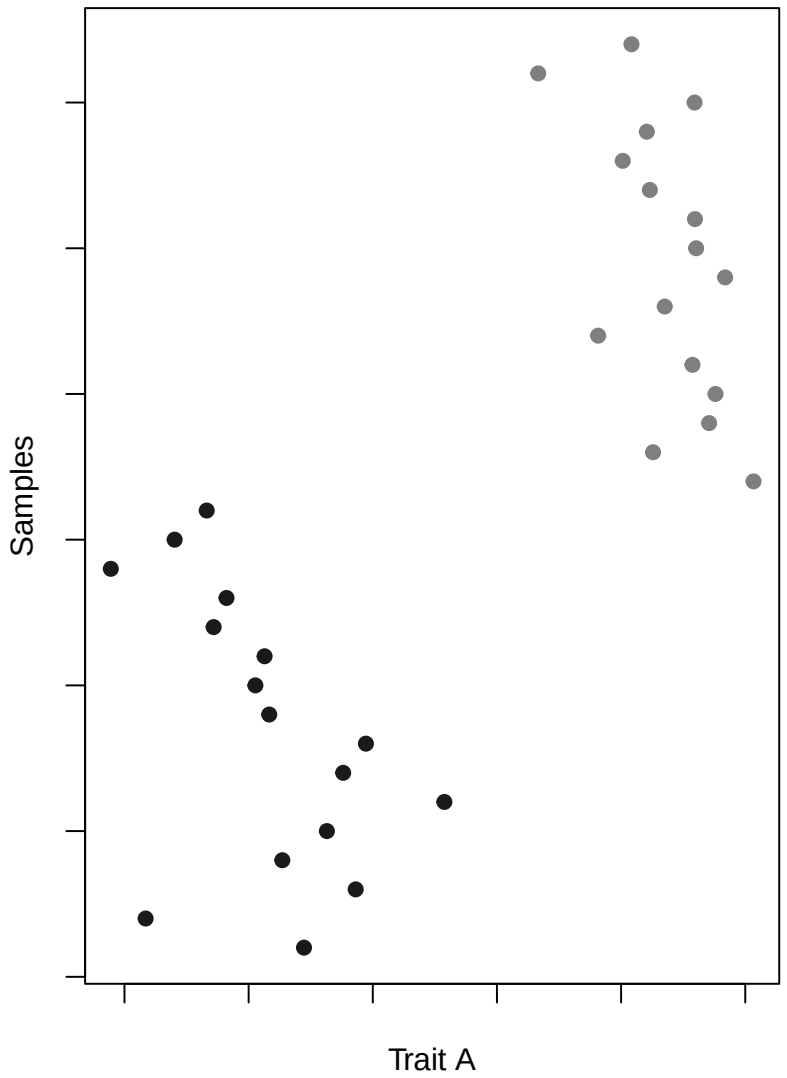
(Felsenstein, 1985)



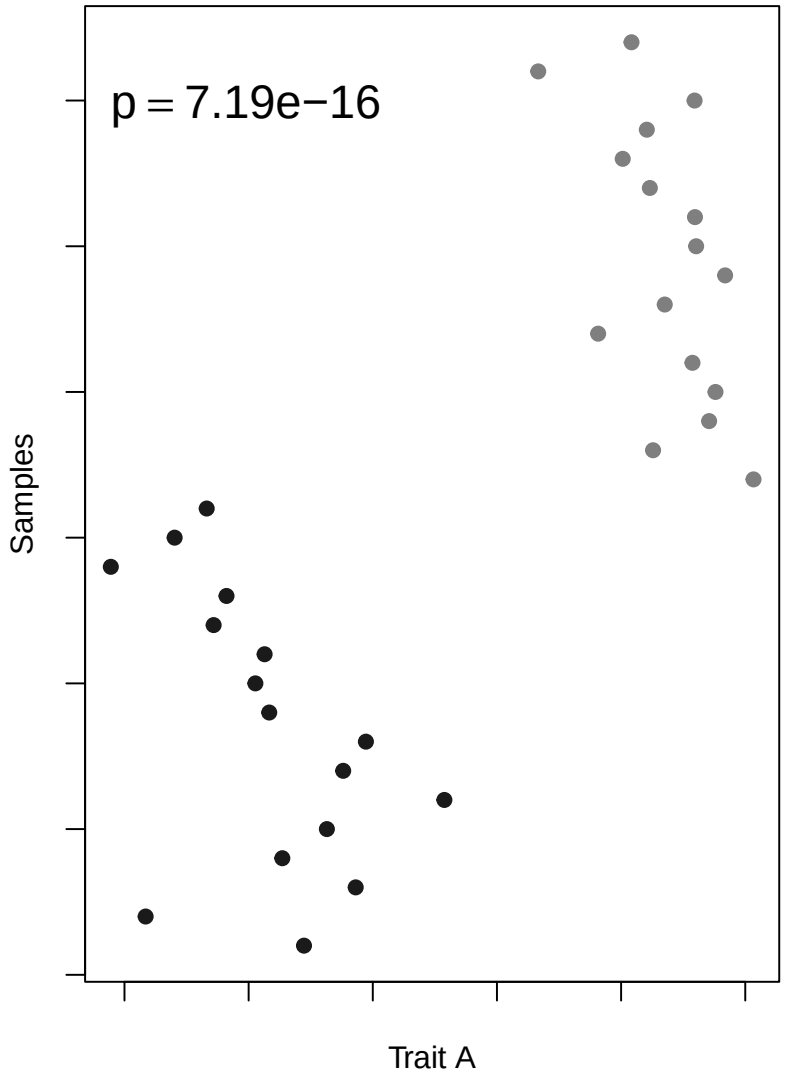




- SDE: $dX_t = \sigma dB_t$
- Covariances: $\mathbb{C}\text{ov}(Y_i, Y_j) = \sigma^2 V_{ij}$
- Distribution: $\mathbf{Y} \sim \mathcal{N}(\mu \mathbf{1}_n, \sigma^2 \mathbf{V})$

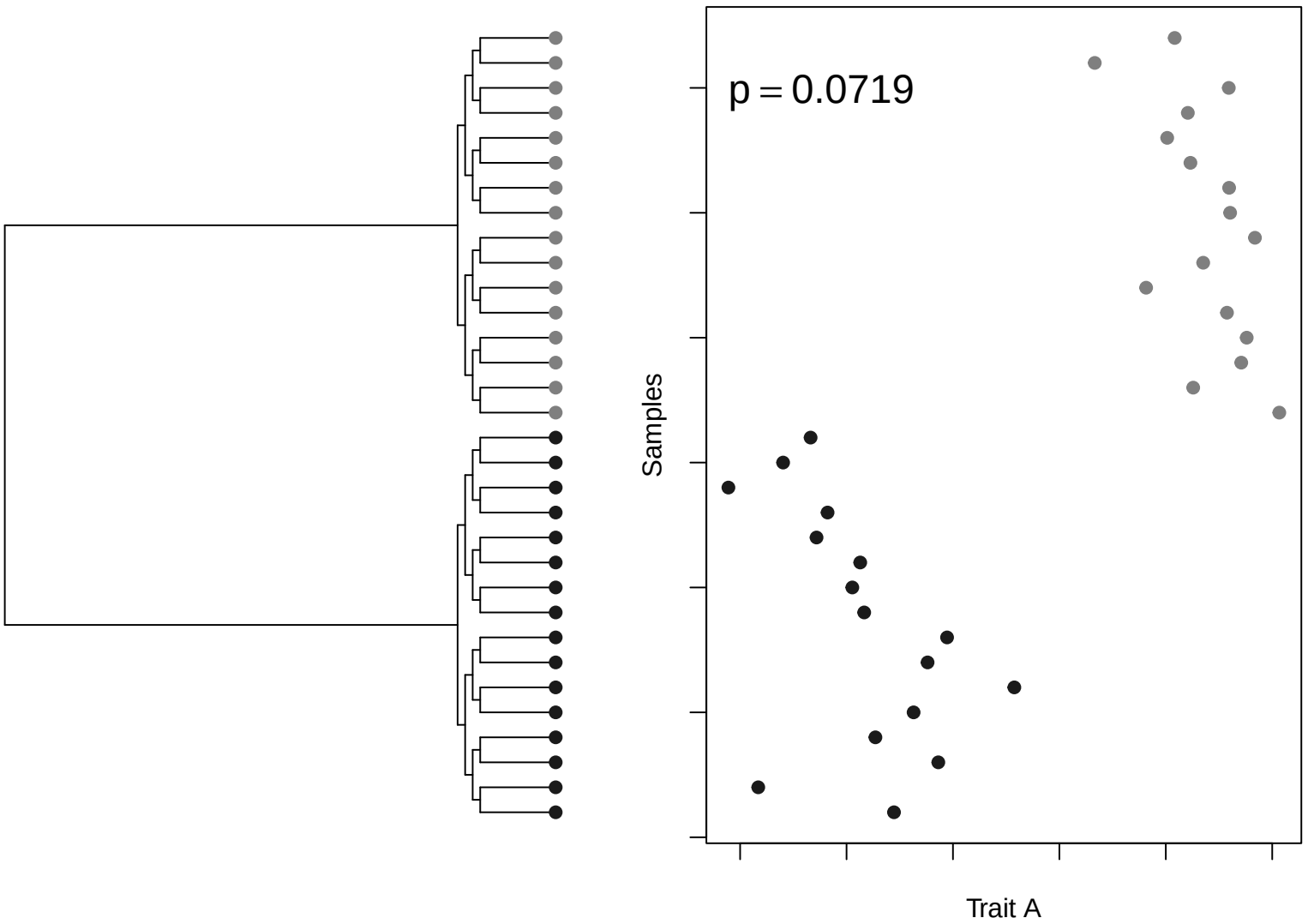
# Phylogenetic ANOVA



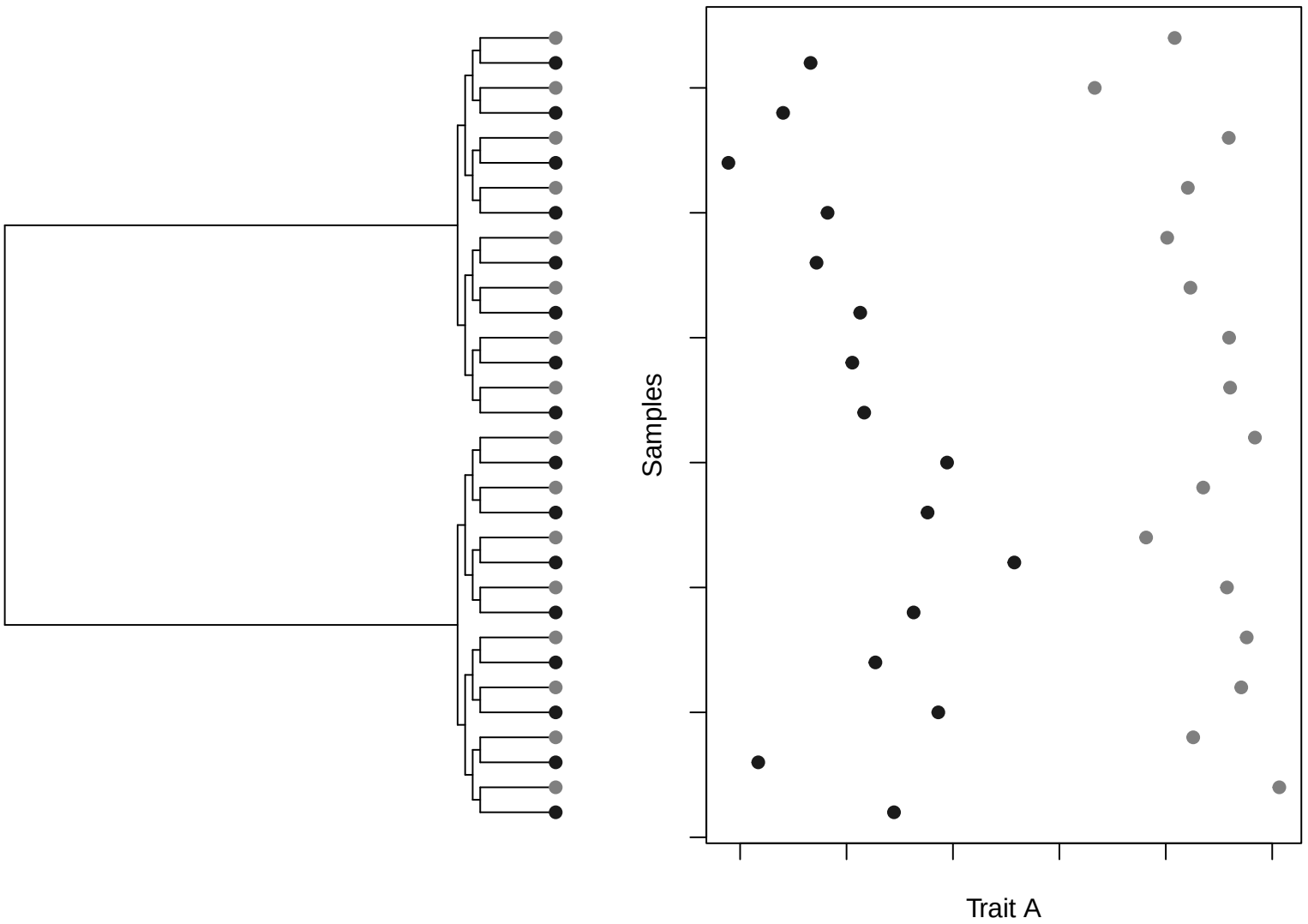
# Phylogenetic ANOVA



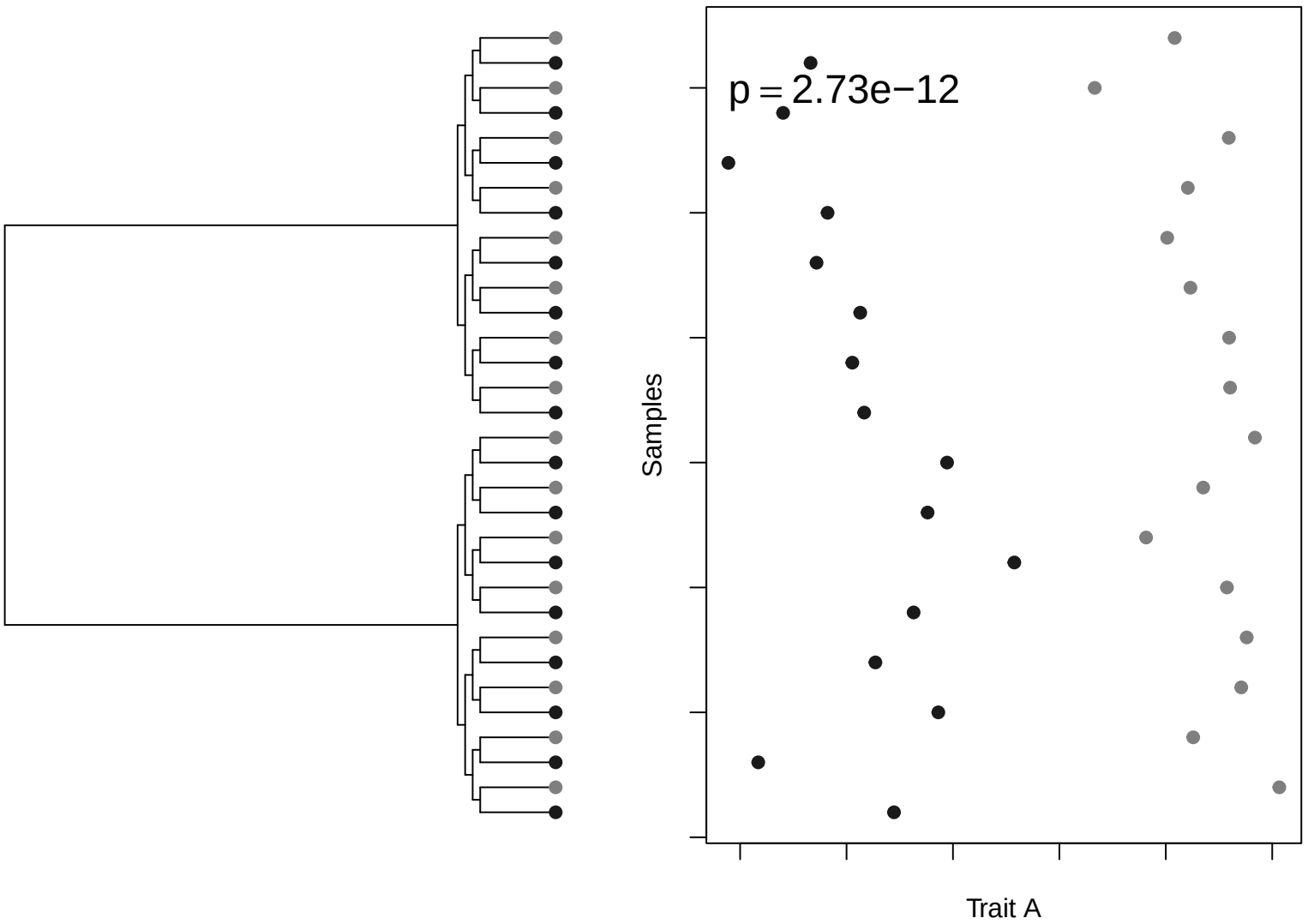
# Phylogenetic ANOVA



# Phylogenetic ANOVA



# Phylogenetic ANOVA



# Phylogenetic ANOVA

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \sigma\mathbf{E}$$

- $\mathbf{Y}$ traits at the tips ($n$)
- $\mathbf{X}$ design matrix ($n \times q$)
- $\boldsymbol{\theta}$ vector of coefficients ($q$)
- $\mathbf{E}$ phylogenetic errors ($n$)

$$\mathbf{E} \sim \mathcal{N}(\mathbf{0}_n, \mathbf{V})$$

$\mathbf{V}$ informed by the tree structure

# Phylogenetic ANOVA

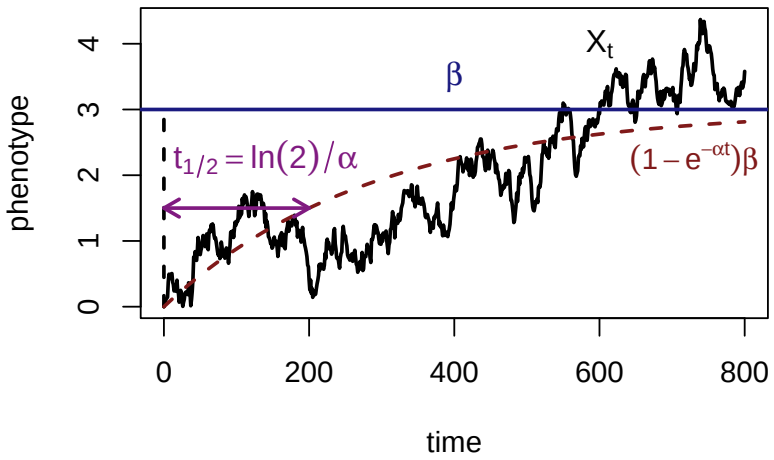
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \sigma\mathbf{E}$$

- $\mathbf{Y}$ traits at the tips ($n$)
- $\mathbf{X}$ design matrix ($n \times q$)
- $\boldsymbol{\theta}$ vector of coefficients ($q$)
- $\mathbf{E}$ phylogenetic errors ($n$)

$$\mathbf{E} \sim \mathcal{N}(\mathbf{0}_n, \mathbf{V}(\phi))$$

$\mathbf{V}(\phi)$ informed by the tree and the trait model

# Ornstein-Uhlenbeck




$$dX_t = \alpha[\beta - X_t]dt + \sigma dB_t$$

Deterministic part:

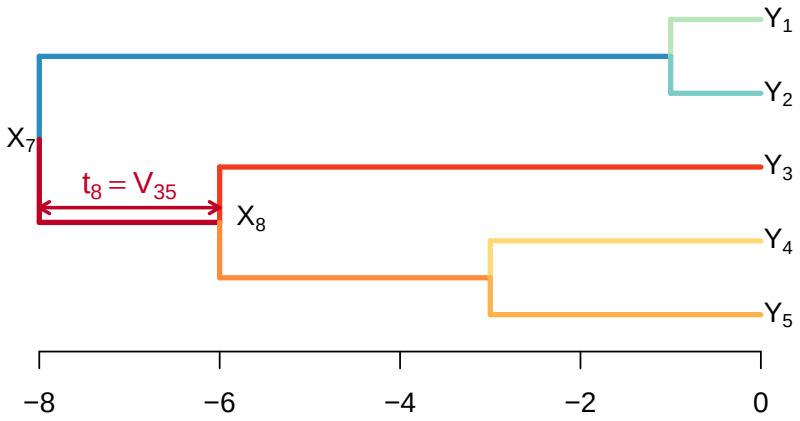
- $\beta$: primary optimum (mechanistically defined).
- $\ln(2)/\alpha$: phylogenetic half live.

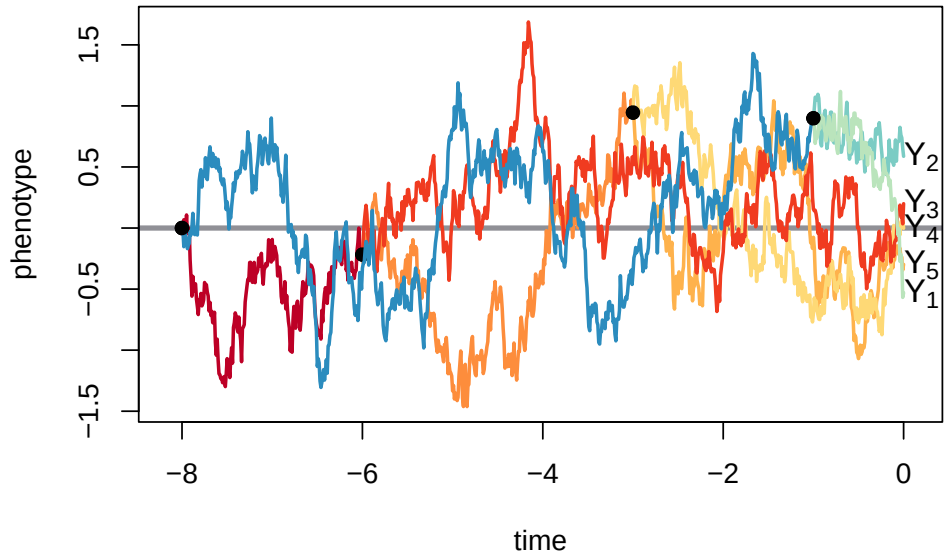
Stochastic part:

- $X_t$: trait value (actual optimum).
- $\sigma dB(t)$: Brownian fluctuations.

# Ornstein-Uhlenbeck on a Tree

(Hansen, 1997)

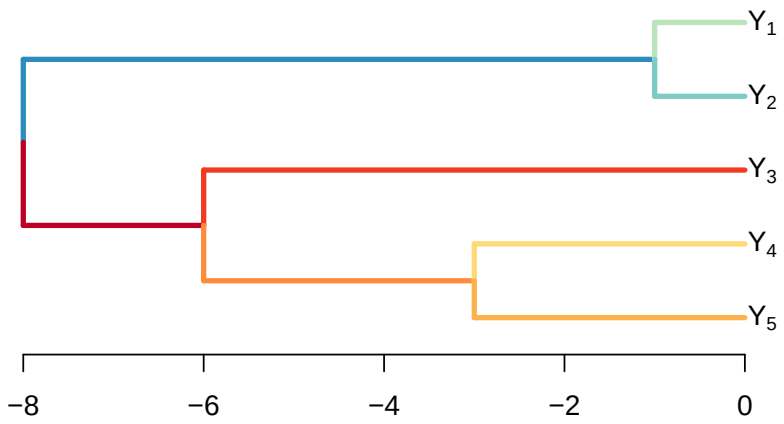


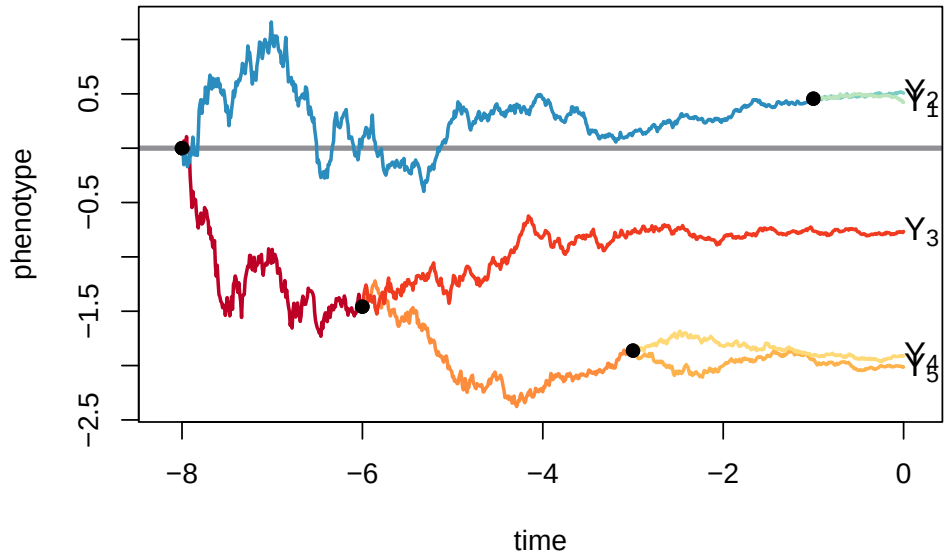




- SDE: $dX_t = \alpha[\beta - X_t]dt + \sigma dB_t$
- Covariances: $\mathbb{C}\text{ov}\,[Y_i; Y_j] = \frac{\sigma^2}{2\alpha} e^{-\alpha(V_i+V_j)}(e^{2\alpha V_{ij}} - 1)$
- Bounded variance $\gamma^2 = \frac{\sigma^2}{2\alpha}$
- Stationary state, Stabilizing selection

# Early Burst

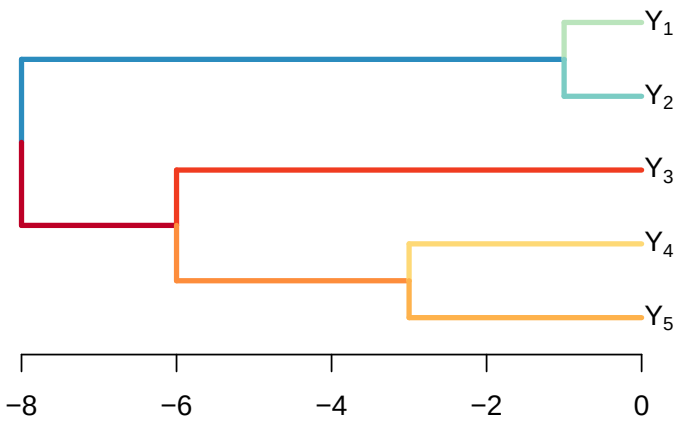
(Harmon et al., 2010)

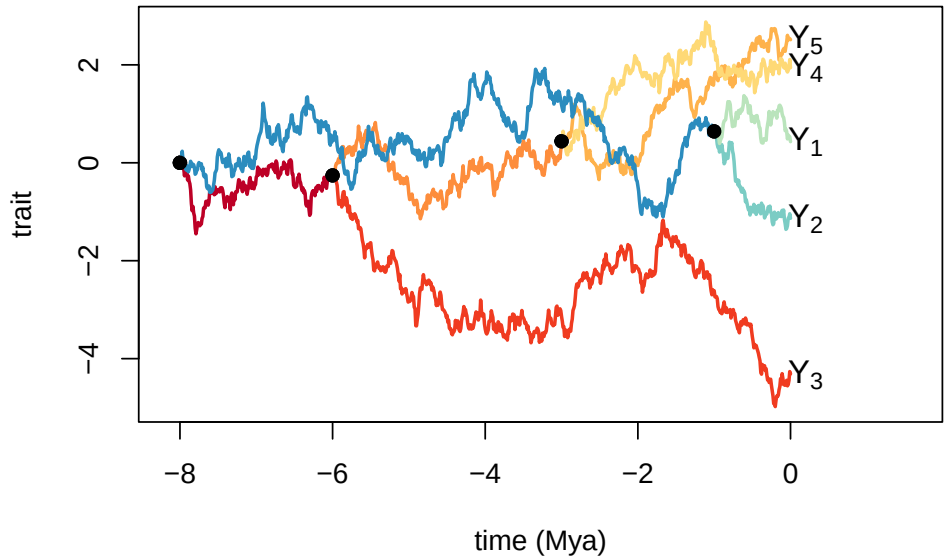



- SDE: $dX_t = \sigma_0 e^{rt/2} dB_t$
- Covariances: $\mathbb{C}\mathrm{ov}\left[Y_i; Y_j\right] = \sigma_0^2 \frac{e^{rV_{ij}} - 1}{r}$

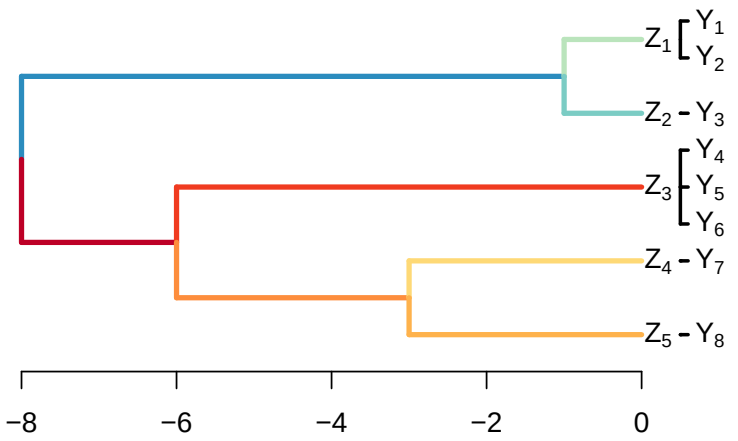
# Multiple Observations

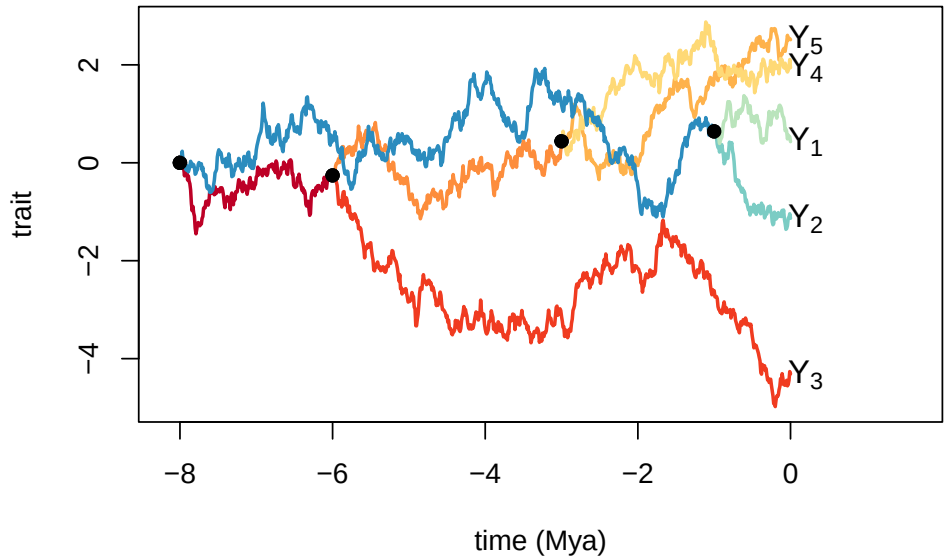



- Covariances: $\mathbb{C}\text{ov}(Y_i, Y_j) = \sigma^2 V_{ij}$
- Distribution: $\mathbf{Y} \sim \mathcal{N}(\mu \mathbf{1}_n, \sigma^2 \mathbf{V})$

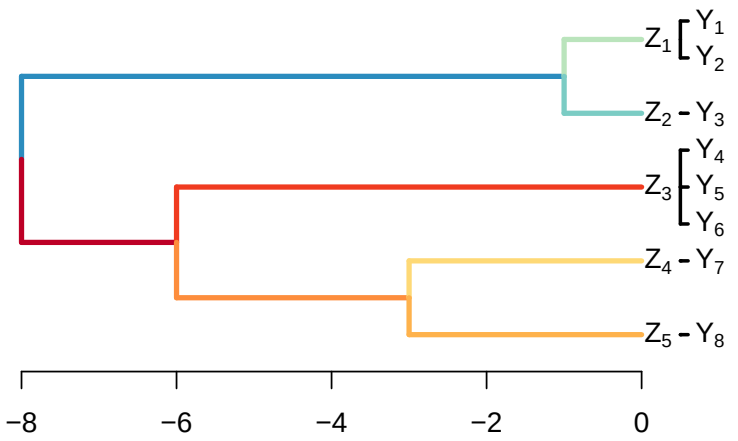
# Multiple Observations

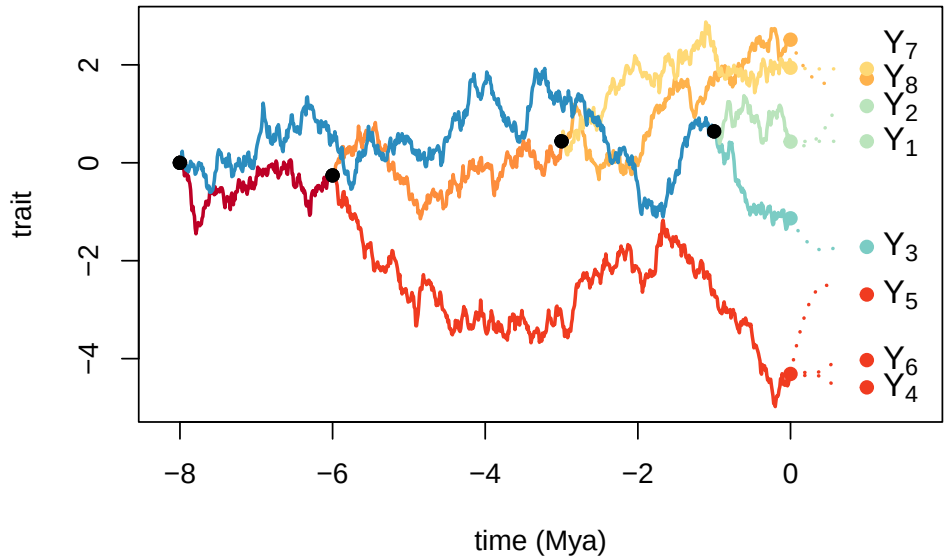



- Covariances: $\mathbb{C}\text{ov}(Y_i, Y_j) = \sigma^2 V_{ij}$
- Distribution: $\mathbf{Y} \sim \mathcal{N}(\mu \mathbf{1}_n, \sigma^2 \mathbf{V})$

# Multiple Observations




- Observation: $Y_1|Z_1 \sim \mathcal{N}(Z_1, \sigma_e^2)$
- Covariances: $\mathbb{C}\text{ov}(Y_i, Y_j) = \sigma^2 V_{ij} + \sigma_e^2 \delta_{ij}$
- Distribution: $\mathbf{Y} \sim \mathcal{N}(\mu \mathbf{1}_n, \sigma^2 \mathbf{V} + \sigma_e^2 \mathbf{I}_n)$

# Phylogenetic ANOVA

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \sigma\mathbf{E} \quad \mathbf{E} \sim \mathcal{N}(\mathbf{0}_n, \mathbf{V}(\phi))$$

BM: $V_{ij} = t_{ij}$

errors: $V(\phi)_{ij} = t_{ij} + \phi h$

EB: $V(r)_{ij} = \dfrac{e^{rV_{ij}} - 1}{r}$

OU: $V(\alpha)_{ij} = e^{-\alpha(V_i + V_j)} \dfrac{e^{2\alpha V_{ij}} - 1}{2\alpha}$

# Phylogenetic ANOVA

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \sigma\mathbf{E} \quad \mathbf{E} \sim \mathcal{N}(\mathbf{0}_n, \mathbf{V}(\phi))$$

BM: $V_{ij} = t_{ij}$ errors: $V(\phi)_{ij} = t_{ij} + \phi h$

EB: $V(r)_{ij} = \frac{e^{rV_{ij}} - 1}{r}$ OU: $V(\alpha)_{ij} = e^{-\alpha(V_i + V_j)} \frac{e^{2\alpha V_{ij}} - 1}{2\alpha}$

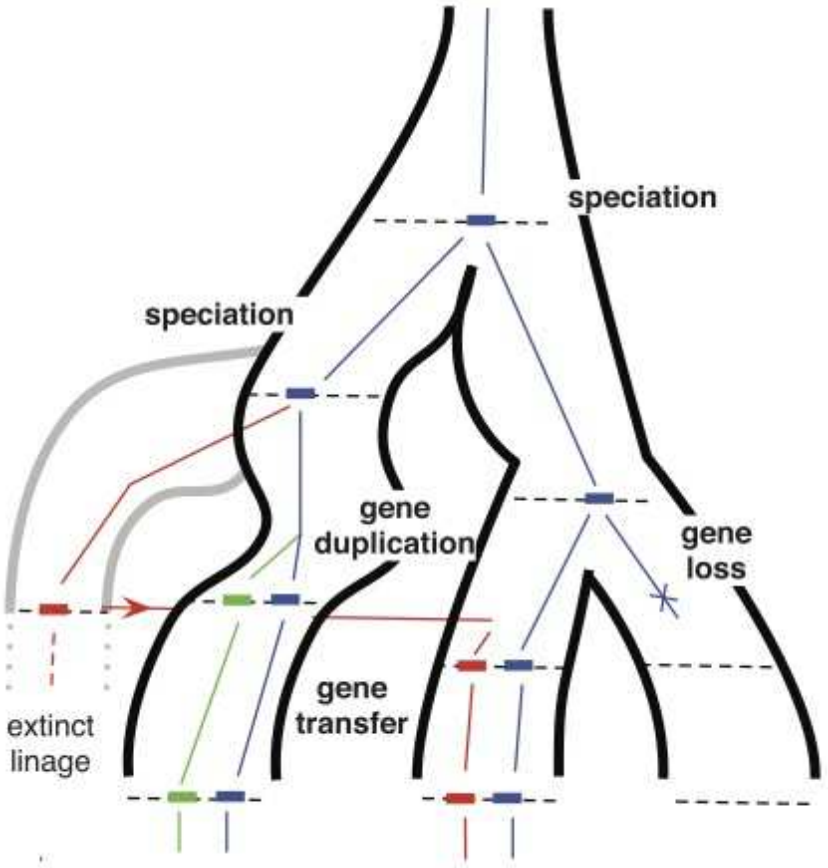
Estimators:

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}^T \mathbf{V}(\hat{\phi})^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}(\hat{\phi})^{-1} \mathbf{Y}$$

$$\hat{\sigma}^2 = \frac{1}{n-p} (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}})^T \mathbf{V}(\hat{\phi})^{-1} (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}}) = \frac{1}{n-p} \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}}\|^2_{\mathbf{V}(\hat{\phi})^{-1}}$$

$\hat{\phi}$ numerical maximization

# Orthologous Genes




(Szöllősi et al., 2015)

- Find Orthologous Genes
  Inparanoid, MaRiO, OrthoFinder, ...

  see Tekaia (2016)
- No reference genome
- Keep only one-to-one
- **Length matters**

# Toy Example for the OG Matrix

| **Species A** | Asample1 | Asample2 |
|---|---|---|
| Acontig1 | 1 | 2 |
| Acontig2 | 7 | 8 |
| Acontig3 | 13 | 14 |

| **Species B** | Bsample1 | Bsample2 |
|---|---|---|
| Bcontig1 | 3 | 4 |
| Bcontig2 | 9 | 0 |
| Bcontig3 | 15 | 16 |

| **Species C** | Csample1 | Csample2 |
|---|---|---|
| Ccontig1 | 5 | 6 |
| Ccontig2 | 11 | 12 |
| Ccontig3 | 17 | 18 |

**OG 1**:
{*Acontig*1, *Bcontig*1, *Ccontig*1}
**OG 2**:
{*Acontig*2, *Acontig*3, *Bcontig*2, *Ccontig*2}
**OG 3**:
{*Bcontig*3, *Ccontig*3}

| | Asample1 | Asample2 | Bsample1 | Bsample2 | Csample1 | Csample2 |
|---|---|---|---|---|---|---|
| OG 1 | 1 | 2 | 3 | 4 | 5 | 6 |
| OG 2 | 7+13 | 8+14 | 9 | 0 | 11 | 12 |
| OG 3 | ∅ | ∅ | 15 | 16 | 17 | 18 |

# DESeq2 with Lengths (Love et al., 2014)

$$Y_{gi} \sim \text{NB}(\mu_{gi}, \alpha_g)$$
$$\mu_{gi} = c_{gi} q_{gi}$$
$$\log_2(q_{gi}) = \mathbf{X}_{i.} \boldsymbol{\theta}_g$$

Normalisation factor: $c_{gi}$ depends on OG length $\ell_{gi}$.

$$c_{gi} = \frac{s_i \ell_{gi}}{\exp(\frac{1}{n} \sum_{i=1}^{n} \log(s_i \ell_{gi}))}$$

(from the DESeq2 tutorial)

Used in: Torres-Oliva et al. (2016)

# Normalization with Lengths

CPM:

$$\text{CPM}_{gi} = \frac{Y_{gi}}{M_i/10^6} \qquad M_i = \sum_g Y_{gi} m_i$$

RPKM: (Mortazavi et al., 2008)

$$\text{RPKM}_{gi} = \frac{Y_{gi}}{M_i/10^6 \times \ell_{gi}/10^3}$$

TPM: (Wagner et al., 2012)

$$\text{TPM}_{gi} = \frac{Y_{gi}/\ell_{gi}}{\sum_g Y_{gi}/\ell_{gi}/10^6}$$

TPM over genes sum to a constant (Musser and Wagner, 2015)

# Normalization with lengths

In the inter-species litterature:

- $\log_2(\text{CPM})$:
  Blake et al. (2018)
- $\log_2(\text{RPKM})$:
  Mortazavi et al. (2008); Brawand et al. (2011); Catalán et al. (2019)
- $\log_{10}(\text{TPM})$:
  Chen et al. (2019)
- $\sqrt{\text{TPM}}$:
  Musser and Wagner (2015); Stern and Crandall (2018)

# `limma` with `duplicateCorrelation` (Smyth et al., 2005)

$$\tilde{\mathbf{Y}}_g = \mathbf{X}\boldsymbol{\theta}_g + \mathbf{E}_g \qquad \mathbf{E}_g \sim \mathcal{N}\left(\mathbf{0}, \sigma_g^2 \mathbf{C}\right)$$

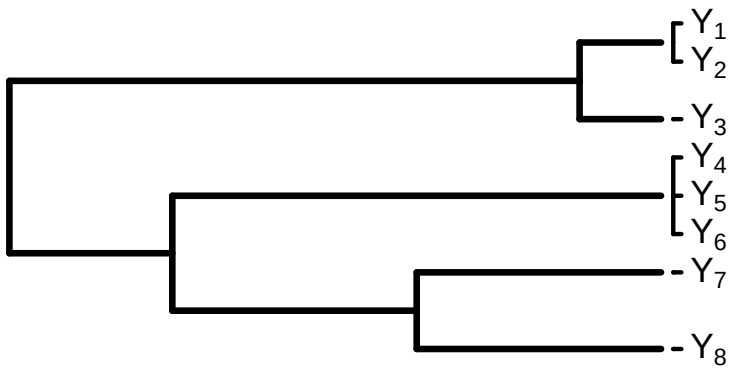
- $\tilde{\mathbf{Y}}_g$: normalized data for gene $g$ $(n \times 1)$
- $\sigma_g^2$ : (moderated) gene specific variance
- $\mathbf{C}$ : (shared) correlation matrix

# `limma` with `duplicateCorrelation` (Smyth et al., 2005)

$$\tilde{\mathbf{Y}}_g = \mathbf{X}\boldsymbol{\theta}_g + \mathbf{E}_g \qquad \mathbf{E}_g \sim \mathcal{N}\left(\mathbf{0}, \sigma_g^2 \mathbf{C}\right)$$

- $\tilde{\mathbf{Y}}_g$: normalized data for gene $g$ $(n \times 1)$
- $\sigma_g^2$: (moderated) gene specific variance
- $\mathbf{C}$ : (shared) correlation matrix

$$\mathbb{C}\text{or}\left[\tilde{Y}_{gi}; \tilde{Y}_{gj}\right] = C_{ij} = \begin{cases} \rho & \text{if } i \text{ and } j \text{ are the same species} \\ 0 & \text{otherwise} \end{cases}$$




$$\mathbf{C} = \begin{array}{c} \\ Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \\ Y_8 \end{array} \begin{array}{c} \begin{array}{cccccccc} Y_1 & Y_2 & Y_3 & Y_4 & Y_5 & Y_6 & Y_7 & Y_8 \end{array} \\ \begin{pmatrix} 1 & \rho & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \rho & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & 1 & \rho & \rho & \cdot & \cdot \\ \cdot & \cdot & \cdot & \rho & 1 & \rho & \cdot & \cdot \\ \cdot & \cdot & \cdot & \rho & \rho & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 \end{pmatrix} \end{array}$$

# `phylolm`: Phylogenetic ANOVA (Ho and Ané, 2014)

$$\tilde{\mathbf{Y}}_g = \mathbf{X}\boldsymbol{\theta}_g + \mathbf{E}_g \qquad \mathbf{E}_g \sim \mathcal{N}(\mathbf{0}, \sigma_g^2 \mathbf{V}(\phi_g))$$

- $\tilde{\mathbf{Y}}_g$: normalized data for gene $g$ $(n \times 1)$
- $\sigma_g^2$: gene specific variance
- $\mathbf{V}(\phi_g)$: tree correlation matrix

BM: $V_{ij} = t_{ij}$ errors: $V(\phi_g)_{ij} = t_{ij} + \phi_g h$

AC/DC: $V(r_g)_{ij} = \dfrac{e^{r_g V_{ij}} - 1}{r_g}$ OU: $V(\alpha_g)_{ij} = e^{-\alpha_g (V_i + V_j)} \dfrac{e^{2\alpha_g V_{ij}} - 1}{2\alpha_g}$

Used in:

Brawand et al. (2011); Rohlfs et al. (2014); Rohlfs and Nielsen (2015); Stern and Crandall (2018); Catalán et al. (2019); Chen et al. (2019)

# Summary: Methods for DE Across Species

| | models counts | requires transformation | shares information across genes | takes phylogeny into account |
|---|---|---|---|---|
| NB-based model (`DESeq`) | yes | no | yes | no |
| Linear model (`limma`) | no | yes | yes | partly |
| Phylogenetic Regression (`phylolm`) | no | yes | no | yes |

Question: Which method performs best ?

# One Species Model

(Robinson and Oshlack, 2010)

$$Y_{gi} \sim \mathrm{NB}(\mu_{gi}, \alpha_g)$$

- $Y_{gi}$ : observed expression — gene $g$, sample $i$
- $\alpha_g$ : dispersion — gene $g$

# One Species Model

(Robinson and Oshlack, 2010)

$$Y_{gi} \sim \text{NB}(\mu_{gi}, \alpha_g)$$

- $Y_{gi}$ : observed expression — gene $g$, sample $i$
- $\alpha_g$ : dispersion — gene $g$
- $\lambda_{gi}$ : true expression — gene $g$, sample $i$
- $M_i$ : sampling depth — sample $i$

$$\mu_{gi} = \frac{\lambda_{gi}}{\sum_{h=1}^{p} \lambda_{hi}} M_i$$

# One Species Model

(Robinson and Oshlack, 2010)

$$Y_{gi} \sim \text{NB}(\mu_{gi}, \alpha_g)$$

- $Y_{gi}$ : observed expression — gene $g$, sample $i$
- $\alpha_g$ : dispersion — gene $g$
- $\lambda_{gi}$ : true expression — gene $g$, sample $i$
- $M_i$ : sampling depth — sample $i$
- $\ell_g$ : gene length — gene $g$

$$\mu_{gi} = \frac{\lambda_{gi}\ell_g}{\sum_{h=1}^{p} \lambda_{hi}\ell_h} M_i.$$

# One Species Model

(Robinson and Oshlack, 2010)

$$Y_{gi} \sim \text{NB}(\mu_{gi}, \alpha_g)$$

- $Y_{gi}$ : observed expression — gene $g$, sample $i$
- $\alpha_g$ : dispersion — gene $g$
- $\lambda_{gi}$ : true expression — gene $g$, sample $i$
- $M_i$ : sampling depth — sample $i$
- $\ell_g$ : gene length — gene $g$
- $S_1$, $S_2$ : groups for differential expression

$$\mu_{gi} = \frac{\lambda_{gi}\ell_g}{\sum_{h=1}^{p} \lambda_{hi}\ell_h} M_i \qquad \lambda_{gi} = \begin{cases} \lambda_{gS_1} & \text{if } i \in S_1 \\ \lambda_{gS_2} & \text{if } i \in S_2 \end{cases}$$

# One Species Model - Simulation (Soneson and Delorenzi, 2013)

$$Y_{gi} \sim \mathsf{NB}(\mu_{gi}, \alpha_g) \quad \mu_{gi} = \frac{\lambda_{gi}\ell_g}{\sum_{h=1}^{p} \lambda_{hi}\ell_h} M_i \quad \lambda_{gi} = \begin{cases} \lambda_{gS_1} & \text{if } i \in S_1 \\ \lambda_{gS_2} & \text{if } i \in S_2 \end{cases}$$

# One Species Model - Simulation (Soneson and Delorenzi, 2013)

$$Y_{gi} \sim \mathsf{NB}(\mu_{gi}, \alpha_g) \quad \mu_{gi} = \frac{\lambda_{gi}\ell_g}{\sum_{h=1}^{p} \lambda_{hi}\ell_h} M_i \quad \lambda_{gi} = \begin{cases} \lambda_{gS_1} & \text{if } i \in S_1 \\ \lambda_{gS_2} & \text{if } i \in S_2 \end{cases}$$

- $\lambda_{gS_2} = \begin{cases} \lambda_{gS_1} & \text{if } g \text{ not DE;} \\ \lambda_{gS_1} \times (e + X_g^e) & \text{if } g \text{ up-regulated in } S_2; \\ \lambda_{gS_1} \times (e + X_g^e)^{-1} & \text{if } g \text{ down-regulated in } S_2. \end{cases}$
- $X_g^e \sim \mathcal{E}(1)$ iid
- $e$ effect size

# One Species Model - Simulation (Soneson and Delorenzi, 2013)

$$Y_{gi} \sim \text{NB}(\mu_{gi}, \alpha_g) \quad \mu_{gi} = \frac{\lambda_{gi}\ell_g}{\sum_{h=1}^{p} \lambda_{hi}\ell_h} M_i \quad \lambda_{gi} = \begin{cases} \lambda_{gS_1} & \text{if } i \in S_1 \\ \lambda_{gS_2} & \text{if } i \in S_2 \end{cases}$$

- $\lambda_{gS_2} = \begin{cases} \lambda_{gS_1} & \text{if } g \text{ not DE;} \\ \lambda_{gS_1} \times (e + X_g^e) & \text{if } g \text{ up-regulated in } S_2; \\ \lambda_{gS_1} \times (e + X_g^e)^{-1} & \text{if } g \text{ down-regulated in } S_2. \end{cases}$
- $X_g^e \sim \mathcal{E}(1)$ iid
- $e$ effect size
- $\ell_g$ known
- $M_i \sim \mathcal{U}(m, M)$
- $\lambda_{gS_1}$, $m$, $M$, $\alpha_g$ calibrated from empirical data.

# One Species Model - From NB to PLN (Chen et al., 2014)

NB Model:

$$Y_{gi} \sim \mathsf{NB}(\mu_{gi}, \alpha_g)$$

# One Species Model - From NB to PLN (Chen et al., 2014)

Poisson-Gamma Model:

$$Z_{gi} \sim \text{Gamma}(1/\alpha_g; \alpha_g \mu_{gi})$$
$$Y_{gi} \mid Z_{gi} \sim \mathcal{P}(Z_{gi})$$

# One Species Model - From NB to PLN (Chen et al., 2014)

Poisson-Gamma Model:

$$Z_{gi} \sim \text{Gamma}(1/\alpha_g; \alpha_g \mu_{gi})$$
$$Y_{gi} \mid Z_{gi} \sim \mathcal{P}(Z_{gi})$$

Variance:

$$\mathbb{V}\text{ar}[Y_{gi}] = \mu_{gi} + \alpha_g \mu_{gi}^2$$

# One Species Model - From NB to PLN (Chen et al., 2014)

Poisson-Gamma Model:

$$Z_{gi} \sim \text{Gamma}(1/\alpha_g; \alpha_g \mu_{gi})$$
$$Y_{gi} \mid Z_{gi} \sim \mathcal{P}(Z_{gi})$$

Variance:

$$\mathbb{V}\text{ar}\left[Y_{gi}\right] = \mu_{gi} + \alpha_g \mu_{gi}^2$$

Coefficient of Variation:

$$\text{CV}(Y_{gi})^2 = \frac{\mathbb{V}\text{ar}\left[Y_{gi}\right]}{\mathbb{E}\left[Y_{gi}\right]^2}$$

# One Species Model - From NB to PLN (Chen et al., 2014)

Poisson-Gamma Model:

$$Z_{gi} \sim \text{Gamma}(1/\alpha_g; \alpha_g \mu_{gi})$$
$$Y_{gi} \mid Z_{gi} \sim \mathcal{P}(Z_{gi})$$

Variance:

$$\mathbb{V}\text{ar}\left[Y_{gi}\right] = \mu_{gi} + \alpha_g \mu_{gi}^2$$

Coefficient of Variation:

$$\text{CV}(Y_{gi})^2 = \frac{1}{\mathbb{E}\left[Z_{gi}\right]} + \text{CV}(Z_{gi})^2$$

# One Species Model - From NB to PLN (Chen et al., 2014)

## Poisson-Gamma Model:

$$Z_{gi} \sim \text{Gamma}(1/\alpha_g; \alpha_g \mu_{gi})$$
$$Y_{gi} \mid Z_{gi} \sim \mathcal{P}(Z_{gi})$$

### Variance:

$$\mathbb{V}\text{ar}\left[Y_{gi}\right] = \mu_{gi} + \alpha_g \mu_{gi}^2$$

### Coefficient of Variation:

$$\text{CV}(Y_{gi})^2 = \frac{1}{\mathbb{E}\left[Z_{gi}\right]} + \text{CV}(Z_{gi})^2$$

## Poisson-Log-Normal Model:

$$Z_{gi} \sim \text{Log-Normal}(m_{gi}; \sigma_g^2)$$
$$Y_{gi} \mid Z_{gi} \sim \mathcal{P}(Z_{gi})\,.$$

# One Species Model - From NB to PLN (Chen et al., 2014)

## Poisson-Gamma Model:

$$Z_{gi} \sim \text{Gamma}(1/\alpha_g; \alpha_g \mu_{gi})$$
$$Y_{gi} \mid Z_{gi} \sim \mathcal{P}(Z_{gi})$$

### Variance:

$$\mathbb{V}\text{ar}\left[Y_{gi}\right] = \mu_{gi} + \alpha_g \mu_{gi}^2$$

### Coefficient of Variation:

$$\text{CV}(Y_{gi})^2 = \frac{1}{\mathbb{E}\left[Z_{gi}\right]} + \text{CV}(Z_{gi})^2$$

## Poisson-Log-Normal Model:

$$Z_{gi} \sim \text{Log-Normal}(m_{gi}; \sigma_g^2)$$
$$Y_{gi} \mid Z_{gi} \sim \mathcal{P}(Z_{gi})\,.$$

### Variance:

$$\mathbb{V}\text{ar}\left[Y_{gi}\right] = \mathbb{E}\left[Z_{gi}\right] + (e^{\sigma_g^2} - 1)\mathbb{E}\left[Z_{gi}\right]^2$$

# One Species Model - From NB to PLN (Chen et al., 2014)

## Poisson-Gamma Model:

$$Z_{gi} \sim \text{Gamma}(1/\alpha_g; \alpha_g \mu_{gi})$$
$$Y_{gi} \mid Z_{gi} \sim \mathcal{P}(Z_{gi})$$

### Variance:

$$\mathbb{V}\text{ar}\left[Y_{gi}\right] = \mu_{gi} + \alpha_g \mu_{gi}^2$$

### Coefficient of Variation:

$$\text{CV}(Y_{gi})^2 = \frac{1}{\mathbb{E}\left[Z_{gi}\right]} + \text{CV}(Z_{gi})^2$$

## Poisson-Log-Normal Model:

$$Z_{gi} \sim \text{Log-Normal}(m_{gi}; \sigma_g^2)$$
$$Y_{gi} \mid Z_{gi} \sim \mathcal{P}(Z_{gi})\,.$$

### Variance:

$$\mathbb{V}\text{ar}\left[Y_{gi}\right] = \mathbb{E}\left[Z_{gi}\right] + (e^{\sigma_g^2} - 1)\mathbb{E}\left[Z_{gi}\right]^2$$

### Coefficient of Variation:

$$\text{CV}(Y_{gi})^2 = \frac{1}{\mathbb{E}\left[Z_{gi}\right]} + \text{CV}(Z_{gi})^2$$

# One Species Model - From NB to PLN - Moments

## Poisson-Gamma Model:

$$\begin{aligned} Z_{gi} &\sim \text{Gamma}(1/\alpha_g; \alpha_g \mu_{gi}) \\ Y_{gi} \mid Z_{gi} &\sim \mathcal{P}(Z_{gi}) \end{aligned}$$

## Poisson-Log-Normal Model:

$$\begin{aligned} Z_{gi} &\sim \text{Log-Normal}(m_{gi}; \sigma_g^2) \\ Y_{gi} \mid Z_{gi} &\sim \mathcal{P}(Z_{gi}) \,. \end{aligned}$$

## Matching Moments:

$$\begin{cases} \sigma_g^2 = \log(1 + \alpha_g) \\ m_{gi} = \log(\mu_{gi}) - \dfrac{1}{2} \log(1 + \alpha_g). \end{cases}$$

# One Species Model - From NB to PLN - Moments

## Poisson-Gamma Model:

$$\begin{aligned} Z_{gi} &\sim \text{Gamma}(1/\alpha_g; \alpha_g \mu_{gi}) \\ Y_{gi} \mid Z_{gi} &\sim \mathcal{P}(Z_{gi}) \end{aligned}$$

## Poisson-Log-Normal Model:

$$\begin{aligned} Z_{gi} &\sim \text{Log-Normal}(m_{gi}; \sigma_g^2) \\ Y_{gi} \mid Z_{gi} &\sim \mathcal{P}(Z_{gi}) \,. \end{aligned}$$

## Matching Moments:

$$\begin{cases} \sigma_g^2 = \log(1 + \alpha_g) \\ m_{gi} = \log(\mu_{gi}) - \dfrac{1}{2} \log(1 + \alpha_g). \end{cases}$$

## Realistic Simulations:

Use the same parameters as the NB model.

# Inter-Species Model: pPLN

## Poisson-Gamma Model:

$$Z_{gi} \sim \text{Gamma}(1/\alpha_g; \alpha_g \mu_{gi})$$
$$Y_{gi} \mid Z_{gi} \sim \mathcal{P}(Z_{gi})$$

## Poisson-Log-Normal Model:

$$Z_{gi} \sim \text{Log-Normal}(m_{gi}; \sigma_g^2)$$
$$Y_{gi} \mid Z_{gi} \sim \mathcal{P}(Z_{gi})\,.$$

# Inter-Species Model: pPLN

Poisson-Gamma Model:

$$Z_{gi} \sim \text{Gamma}(1/\alpha_g; \alpha_g \mu_{gi})$$
$$Y_{gi} \mid Z_{gi} \sim \mathcal{P}(Z_{gi})$$

**Phylogenetic** PLN Model:

$$\mathbf{Z}_g \sim \mathcal{N}(\mathbf{m}_g, \sigma_g^2 \mathbf{V}(\phi))$$
$$Y_{gi} \mid Z_{gi} \sim \mathcal{P}(\exp(Z_{gi}))\,.$$

# Inter-Species Model: pPLN

Poisson-Gamma Model:

$$Z_{gi} \sim \text{Gamma}(1/\alpha_g; \alpha_g \mu_{gi})$$
$$Y_{gi} \mid Z_{gi} \sim \mathcal{P}(Z_{gi})$$

**Phylogenetic** PLN Model:

$$\mathbf{Z}_g \sim \mathcal{N}(\mathbf{m}_g, \sigma_g^2 \mathbf{V}(\phi))$$
$$Y_{gi} \mid Z_{gi} \sim \mathcal{P}(\exp(Z_{gi}))\,.$$

Matching Moments:

$$\begin{cases} \sigma_g^2 T = \log(1 + \alpha_g) \\ m_{gi} = \log(\mu_{gi}) - \dfrac{1}{2}\log(1 + \alpha_g). \end{cases}$$

Constraint: constant diagonal

$$[\mathbf{V}(\phi)]_{ii} \equiv T$$

$\rightarrow$ true for an **ultrametric** tree.

# Inter Species Model - Simulation

$$\mathbf{Z}_g \sim \mathcal{N}\left(\mathbf{m}_g, \sigma_g^2 \mathbf{V}(\phi)\right)$$
$$Y_{gi} \mid Z_{gi} \sim \mathcal{P}(\exp(Z_{gi}))\,.$$

# Inter Species Model - Simulation

$$\mathbf{Z}_g \sim \mathcal{N}\left(\mathbf{m}_g, \sigma_g^2 \mathbf{V}(\phi)\right)$$
$$Y_{gi} \mid Z_{gi} \sim \mathcal{P}(\exp(Z_{gi}))\,.$$

- $m_{gi}$, $\sigma_g^2$ chosen to match moments of Soneson and Delorenzi (2013)
- $\ell_{gi}$ known or simulated

# Inter Species Model - Simulation

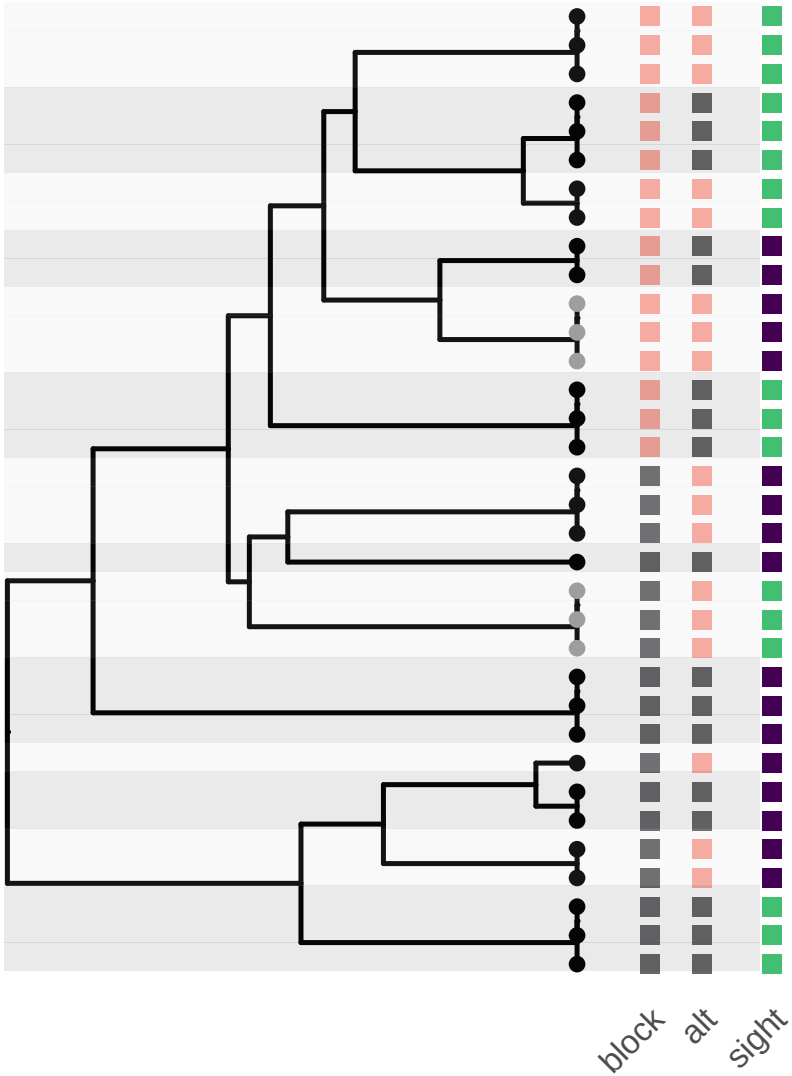
$$\mathbf{Z}_g \sim \mathcal{N}\left(\mathbf{m}_g, \sigma_g^2 \mathbf{V}(\phi)\right)$$
$$Y_{gi} \mid Z_{gi} \sim \mathcal{P}(\exp(Z_{gi}))\,.$$

- $m_{gi}$, $\sigma_g^2$ chosen to match moments of Soneson and Delorenzi (2013)
- $\ell_{gi}$ known or simulated
- $\mathbf{V}(\phi)$ : choose phylogenetic model
  - BM
  - OU
  - with errors

$$t_{1/2} = 50\%$$
$$\frac{s^2}{\sigma_g^2} \in \{0\%, 20\%, 40\%\}$$

# Simulation Setting

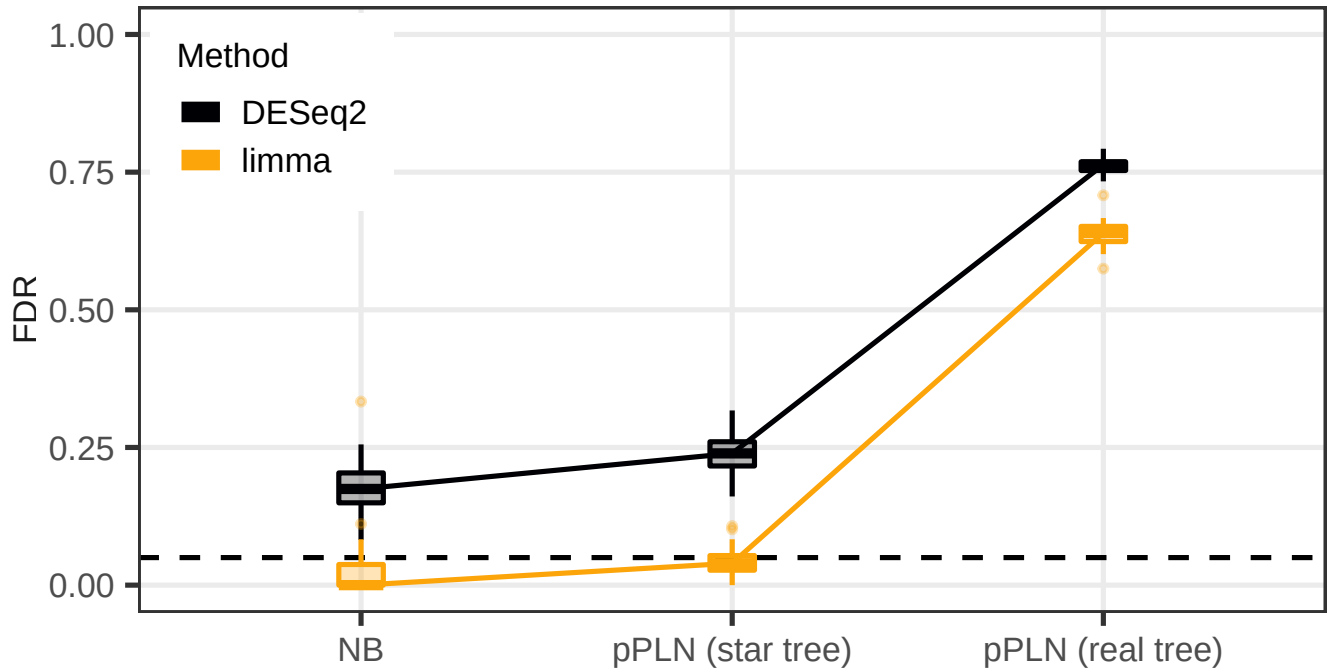
(Stern and Crandall, 2018)



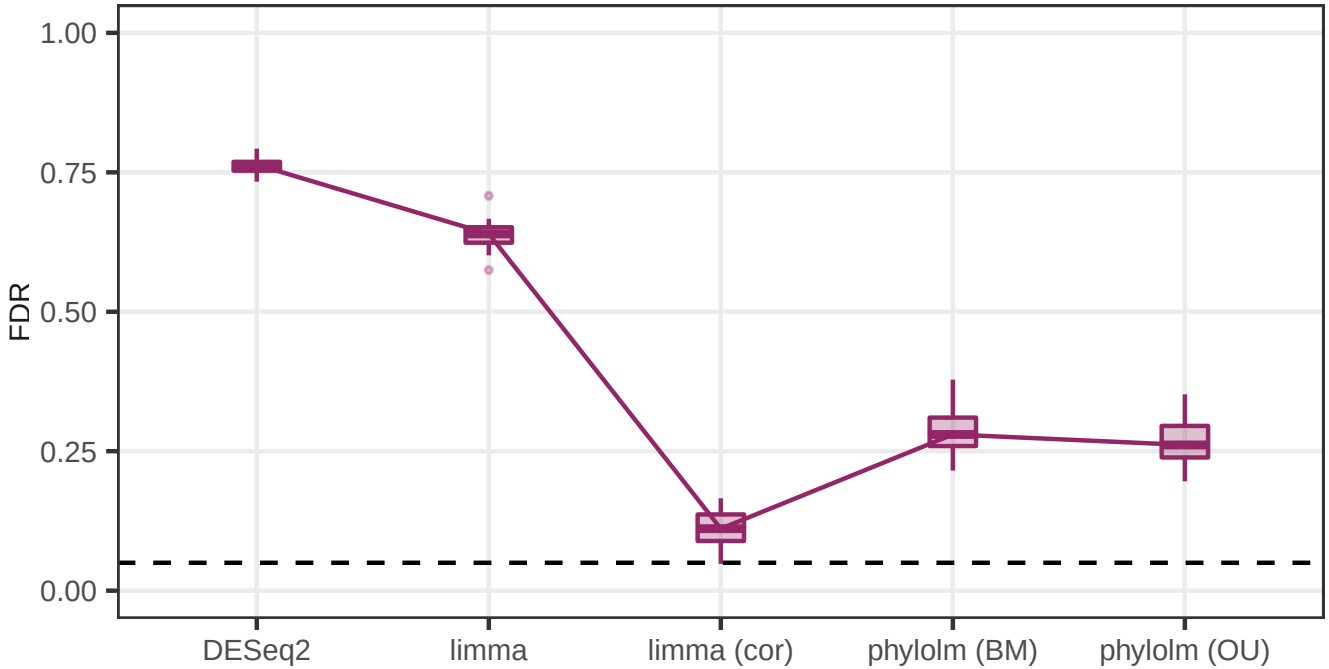

*Orconectes australis*

*Cambarus dubius*

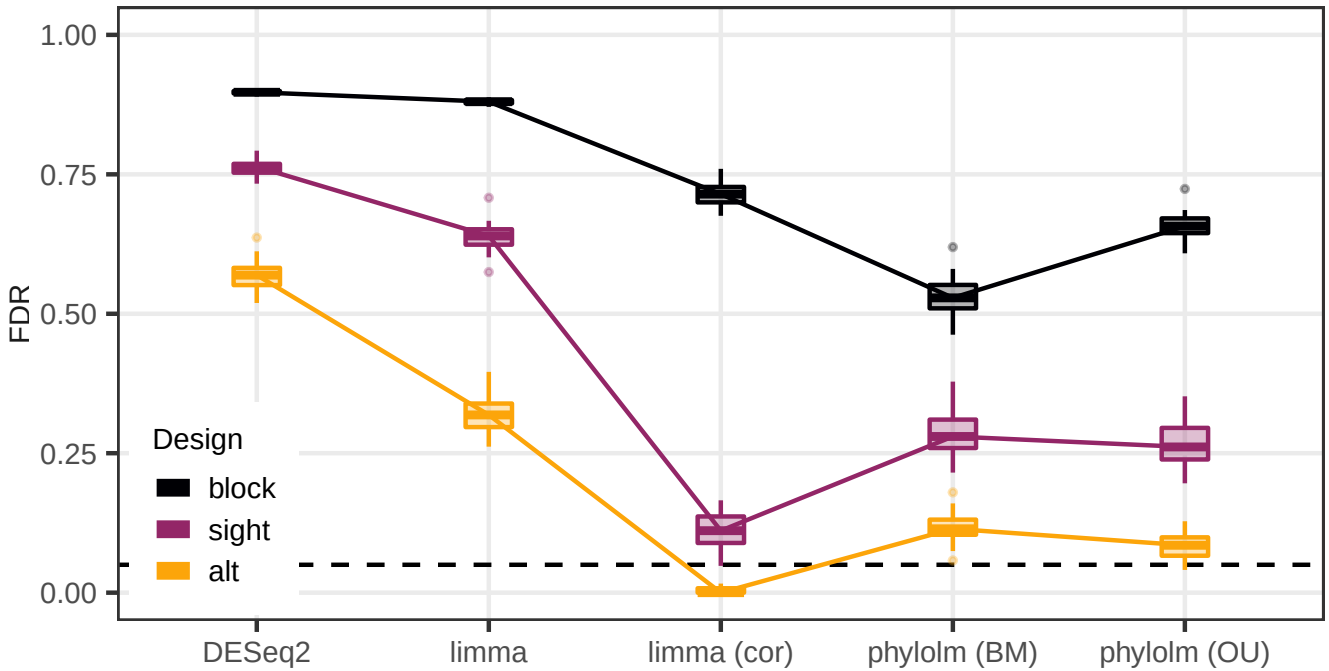
# Tree Correlation Matters



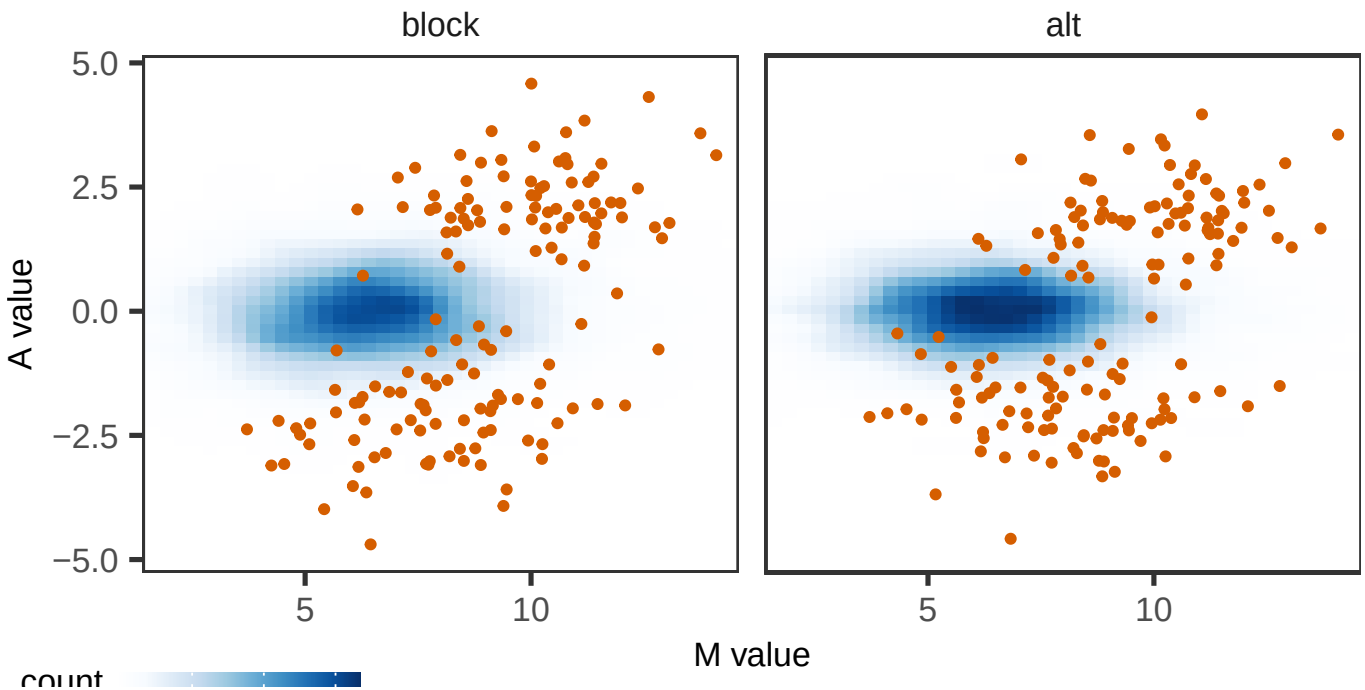
# Tree Correlation and Regularisation Matter



# Tree Group Design Matters



# Tree Group Design Matters



# Differential Analysis Phylogenetic Asymptotic ESS

## Simple BM ANOVA

$$\mathbf{y} = \theta_0 \mathbf{1} + \theta_1 \mathbf{x} + \sigma \mathbf{e}^{BM} \qquad \mathbb{V}\text{ar}\left[\mathbf{e}^{BM}\right] = \mathbf{V}^{\text{tree}} = [t_{ij}]_{i,j}$$

# Differential Analysis Phylogenetic Asymptotic ESS

## Simple BM ANOVA

$$\mathbf{y} = \theta_0 \mathbf{1} + \theta_1 \mathbf{x} + \sigma \mathbf{e}^{BM} \qquad \mathbb{V}\text{ar}\left[\mathbf{e}^{BM}\right] = \mathbf{V}^{\text{tree}} = [t_{ij}]_{i,j}$$

## Estimator Variance:

$$\mathbb{V}\text{ar}\left[\hat{\theta}_1\right] = \sigma^2 (\mathbf{X}^T \mathbf{V}^{\text{tree}\,-1} \mathbf{X})^{-1}_{2,2} \qquad \mathbf{X} = (\mathbf{1}\ \mathbf{x})$$

# Differential Analysis Phylogenetic Asymptotic ESS

## Simple BM ANOVA

$$\mathbf{y} = \theta_0 \mathbf{1} + \theta_1 \mathbf{x} + \sigma \mathbf{e}^{BM} \qquad \mathbb{V}\text{ar}\left[\mathbf{e}^{BM}\right] = \mathbf{V}^{\text{tree}} = [t_{ij}]_{i,j}$$

## Estimator Variance:

$$\mathbb{V}\text{ar}\left[\hat{\theta}_1\right] = \sigma^2 (\mathbf{X}^T \mathbf{V}^{\text{tree}\,-1} \mathbf{X})^{-1}_{2,2} \qquad \mathbf{X} = (\mathbf{1}\ \mathbf{x})$$

## Independent and Balanced Data (star tree):

$$\mathbb{V}\text{ar}\left[\hat{\theta}_1\right] = \sigma^2 \frac{4}{n}$$

# Differential Analysis Phylogenetic Asymptotic ESS

## Simple BM ANOVA

$$\mathbf{y} = \theta_0 \mathbf{1} + \theta_1 \mathbf{x} + \sigma \mathbf{e}^{BM} \qquad \mathbb{V}\text{ar}\left[\mathbf{e}^{BM}\right] = \mathbf{V}^{\text{tree}} = [t_{ij}]_{i,j}$$

## Estimator Variance:

$$\mathbb{V}\text{ar}\left[\hat{\theta}_1\right] = \sigma^2 (\mathbf{X}^T \mathbf{V}^{\text{tree}\,-1} \mathbf{X})^{-1}_{2,2} \qquad \mathbf{X} = (\mathbf{1}\ \mathbf{x})$$

## Independent and Balanced Data (star tree):

$$\mathbb{V}\text{ar}\left[\hat{\theta}_1\right] = \sigma^2 \frac{4}{n}$$

## Normalized dapaESS:

$$\text{dapaESSn}(\mathcal{T}, \mathbf{x}) = \frac{1/(\mathbf{X}^T \mathbf{V}^{\text{tree}\,-1} \mathbf{X})^{-1}_{2,2}}{n/4}$$

# Differential Analysis Phylogenetic Asymptotic ESS

block design:

$$\text{dapaESSn}(\mathcal{T}, \mathbf{x}) = 0.69$$

$\rightarrow$ **harder** than the independent case

sight design:

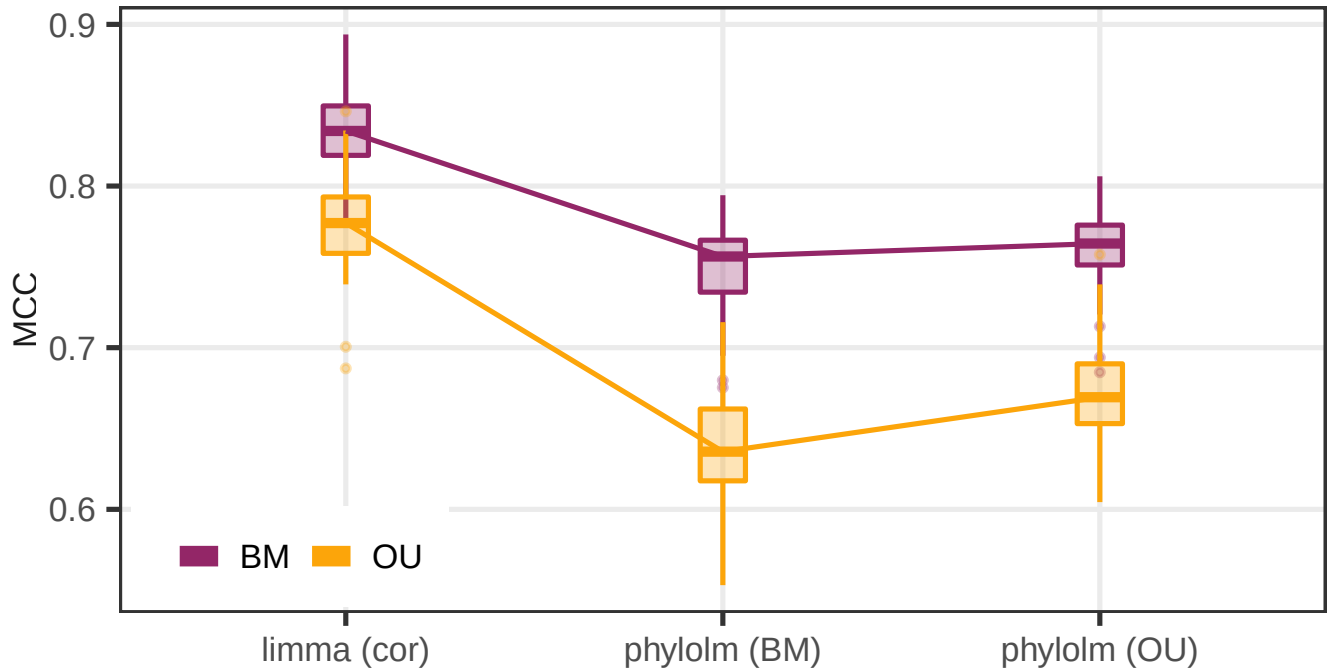
$$\text{dapaESSn}(\mathcal{T}, \mathbf{x}) = 1.4$$

alt design:

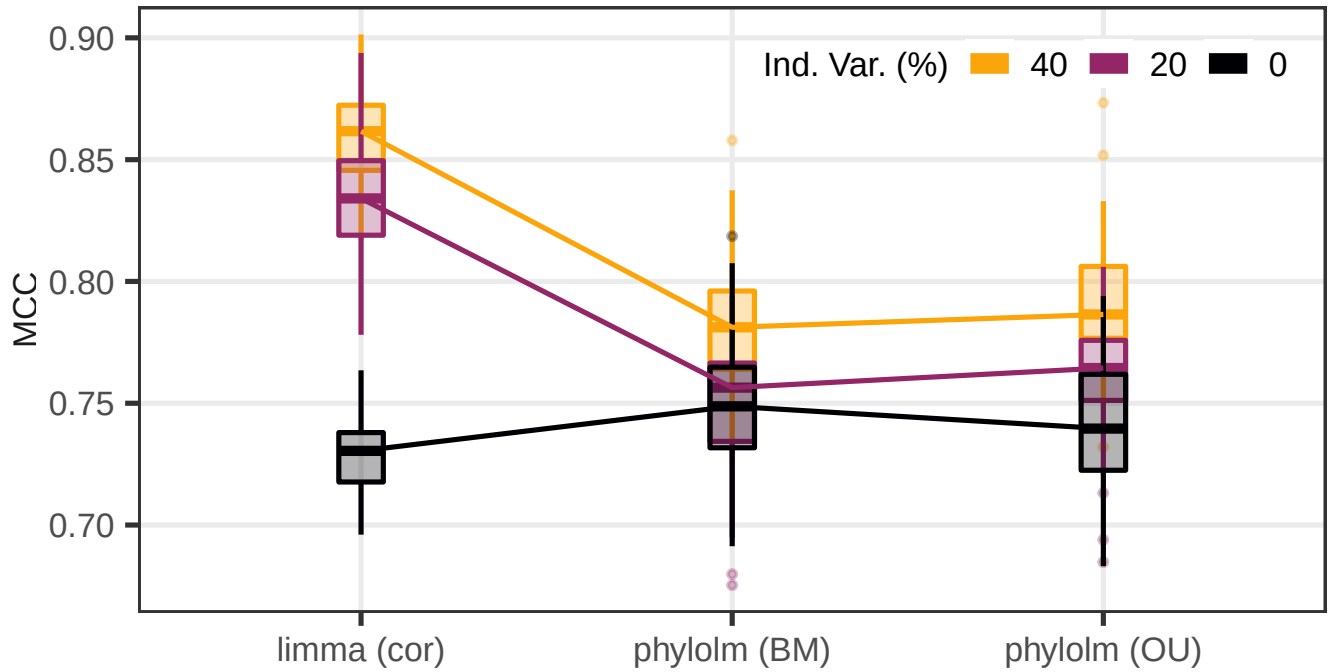
$$\text{dapaESSn}(\mathcal{T}, \mathbf{x}) = 5.1$$

$\rightarrow$ **easier** than the independent case

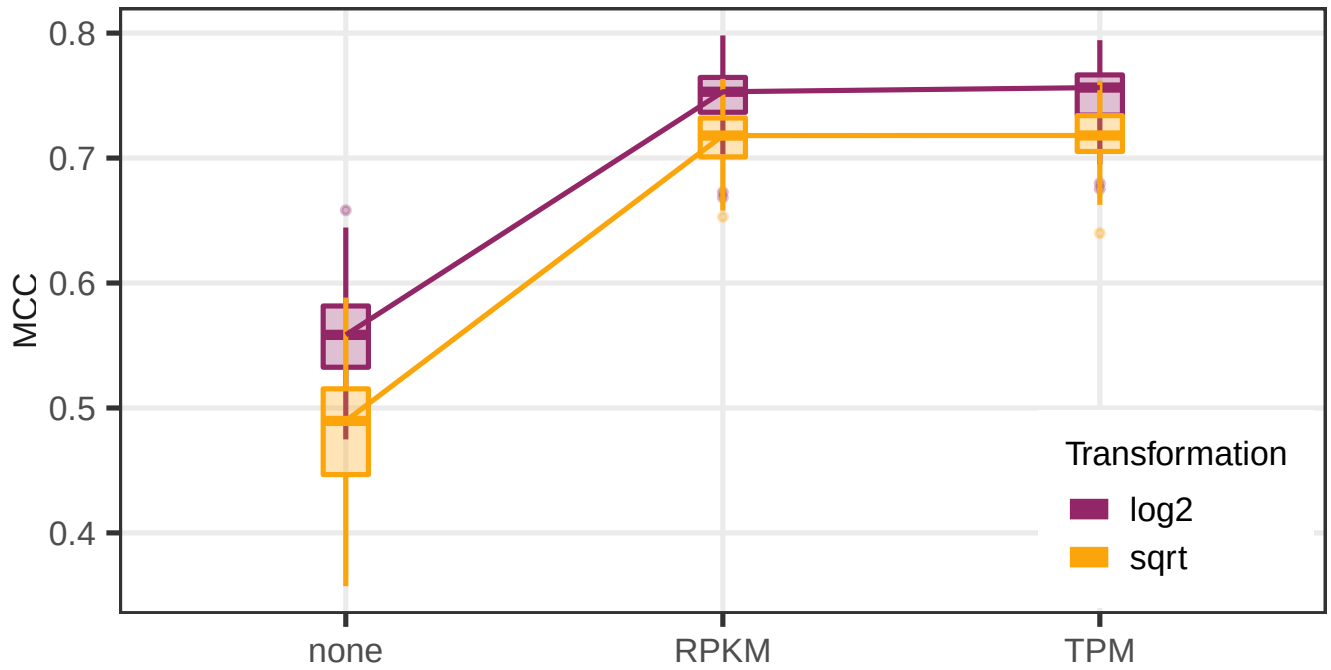
# OU Makes the Signal Weaker



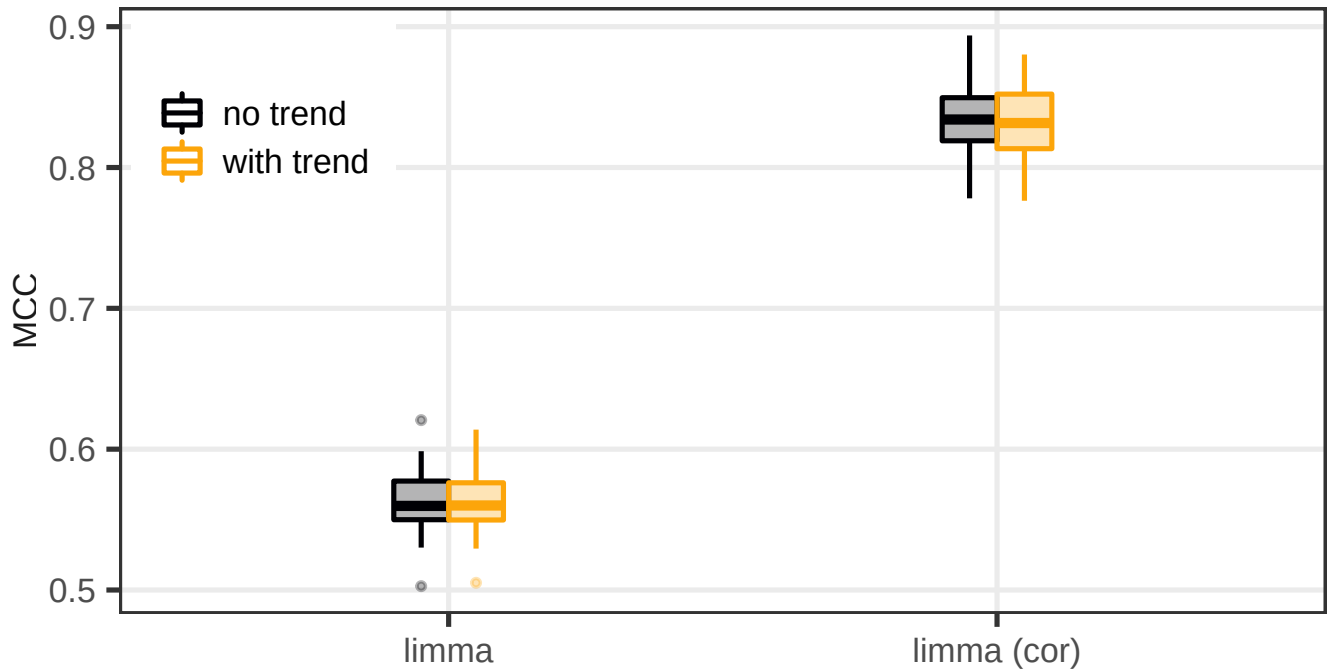
# Intra-Specific Variations



# Normalization



# Trend in eBayes Correction



# Crayfish Data

Stern and Crandall (2018): 93 DE genes
phylolm OU: 17 DE genes
limma cor: 6 DE genes

| Orthogroup | adj. p-value | Uniprot top hit | Protein name |
|---|---|---|---|
| OG0002505 | 2.3e-09 | XYLA ARATH | Xylose isomerase |
| OG0001105 | 4.4e-03 | PIPA DROME | 1-phosphatidylinositol 4,5-bisphosphate phosphodiesterase |
| OG0000233 | 6.2e-03 | RTBS DROME | Probable RNA-directed DNA polymerase from transposon BS |
| OG0002370 | 1.8e-02 | ARRH LOCMI | Arrestin homolog |
| OG0006977 | 2.3e-02 | CSK2B RAT | Casein kinase II subunit beta |
| OG0001281 | 2.9e-02 | OPSD PROCL | Rhodopsin |

Question: Same mechanisms of vision loss in each group ?
$\rightarrow$ analysis by clade.

# Conclusion and Perspectives

Simulation Framework:

- Uses both phylogenetic and RNA-Seq specificities

Inter-Species RNA-Seq Data:

- Group Design matters

A new statistical tool is needed:

- Include phylogenies in RNA-Seq analyses

Bastide P., Soneson C., Lespinet O., Gallopin M. (2022). bioRxiv *Benchmark of Differential Gene Expression Analysis Methods for Inter-species RNA-Seq Data using a Phylogenetic Simulation Framework*.

`compcodeR v1.32`

# Conclusion and Perspectives

Simulation Framework:

- Uses both phylogenetic and RNA-Seq specificities

Inter-Species RNA-Seq Data:

- Group Design matters

A new statistical tool is needed:

- Include phylogenies in RNA-Seq analyses

Bastide P., Soneson C., *Stern D.B.*, Lespinet O., Gallopin M.
In revision.
*A Phylogenetic Framework to Simulate Synthetic Inter-species RNA-Seq Data.*

`compcodeR v1.32`

# Bibliography I


Blake, Thomas, Blischak, et al. 2018. *Genome Biology*. 19:162.

Brawand, Soumillon, Necsulea, et al. 2011. *Nature*. 478:343–348.

Catalán, Briscoe, Höhna. 2019. *Genetics*. 213:581–594.

Chen, Lun, Smyth. 2014. In: Datta, Nettleton, editors, Statistical Analysis of Next Generation Sequencing Data, Cham: Springer International Publishing, pp. 51–74.

Chen, Swofford, Johnson, et al. 2019. *Genome Research*. 29:53–63.

Felsenstein. 1985. *The American Naturalist*. 125:1–15.

Hansen. 1997. *Evolution*. 51:1341.

Harmon, Losos, Davies, et al. 2010. *Evolution*. 64:2385–2396.

Ho, Ané. 2014. *Systematic Biology*. 63:397–408.

Loughman. 2010. *Southeastern Naturalist*. 9:217–230.

Love, Huber, Anders. 2014. *Genome Biology*. 15:550.

Mortazavi, Williams, McCue, et al. 2008. *Nature Methods*. 5:621–628.

Musser, Wagner. 2015. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*. 324:588–604.

Robinson, Oshlack. 2010. *Genome Biology*. 11:R25.

# Bibliography II

Rohlfs, Harrigan, Nielsen. 2014. *Molecular Biology and Evolution*. 31:201–211.

Rohlfs, Nielsen. 2015. *Systematic Biology*. 64:695–708.

Smyth. 2004. *Statistical Applications in Genetics and Molecular Biology*. 3:1–25.

Smyth, Michaud, Scott. 2005. *Bioinformatics*. 21:2067–2075.

Soneson, Delorenzi. 2013. *BMC Bioinformatics*. 14:91.

Stern, Crandall. 2018. *Molecular Biology and Evolution*. 35:2005–2014.

Szöllősi, Tannier, Daubin, et al. 2015. *Systematic Biology*. 64:e42–e62.

Tekaia. 2016. *Genomics Insights*. 9:GEI.S37925.

Torres-Oliva, Almudi, McGregor, et al. 2016. *BMC Genomics*. 17.

Wagner, Kin, Lynch. 2012. *Theory in Biosciences*. 131:281–285.

**Photo Credits**

- *Orconectes australis*: Marshal Hedin, CC BY-SA 2.5
- *Cambarus dubius*: Loughman (2010)

# Thank you for listening

Institut Montpelliérain Alexander Grothendieck

# Appendices